knowledge

MDPI

Article

Epistemology in the Age of Large Language Models

Jennifer Mugleston [1,†], Vuong Hung Truong [1,†], Cindy Kuang [1], Lungile Sibiya [1] and Jihwan Myung [1,2,*]

[1] Graduate Institute of Mind, Brain and Consciousness (GIMBC), Taipei Medical University, New Taipei City 235, Taiwan; d755111002@tmu.edu.tw (J.M.); d755112005@tmu.edu.tw (V.H.T.); m755112004@tmu.edu.tw (C.K.); m755111003@tmu.edu.tw (L.S.)
[2] Graduate Institute of Medical Sciences (GIMS), Taipei Medical University, Taipei 110, Taiwan
* Correspondence: jihwan@tmu.edu.tw
† These authors contributed equally to this work.

Abstract: Epistemology and technology have been working in synergy throughout history. This relationship has culminated in large language models (LLMs). LLMs are rapidly becoming integral parts of our daily lives through smartphones and personal computers, and we are coming to accept the functionality of LLMs as a given. As LLMs become more entrenched in societal functioning, questions have begun to emerge: Are LLMs capable of real understanding? What is knowledge in LLMs? Can knowledge exist independently of a conscious observer? While these questions cannot be answered definitively, we can argue that modern LLMs are more than mere symbol-manipulators and that LLMs in deep neural networks should be considered capable of a form of knowledge, though it may not qualify as justified true belief (JTB) in the traditional definition. This deep neural network design may have endowed LLMs with the capacity for internal representations, basic reasoning, and the performance of seemingly cognitive tasks, possible only through a compressive but generative form of representation that can be best termed as knowledge. In addition, the non-symbolic nature of LLMs renders them incompatible with the criticism posed by Searle's "Chinese room" argument. These insights encourage us to revisit fundamental questions of epistemology in the age of LLMs, which we believe can advance the field.

Keywords: epistemology; large language models (LLMs); knowledge; understanding; Chinese Room

Academic Editor: Jose María Merigo

Received: 15 June 2024
Revised: 25 October 2024
Accepted: 22 January 2025
Published: 1 February 2025

Citation: Mugleston, J.; Truong, V.H.; Kuang, C.; Sibiya, L.; Myung, J. Epistemology in the Age of Large Language Models. *Knowledge* **2025**, *5*, 3. https://doi.org/10.3390/knowledge5010003

1. Introduction

Epistemology is the theory of knowledge. The ancient Greeks had specific terms for various kinds of knowledge. Some of these terms survive in our age: *Historia* refers to witnessed facts or knowledge gained through investigation, and it remains in the word history. *Techne*, the crafty knowledge remembered by the hands for practical skills, is related to technique. *Gnosis*, which refers to mystical knowledge of the divine, is related to cognition, and in neurology, *agnosia* means the inability to recognize. *Doxa* is an opinion or belief that is not scientifically justified or validated as truth. Plato uses this term to contrast with the true knowledge, *episteme*.

Episteme occupies a unique position in the theory of knowledge, signifying a systematic and theoretical understanding akin to scientific knowledge. While natural philosophy evolved into the science of physics, epistemology remained a philosophical inquiry into

how knowledge is defined, acquired, and justified. With the rise of computing, the boundaries of human knowledge have been increasingly challenged. From early contributions of computer simulations in physics to advancements in pattern recognition using perceptrons, and later, breakthroughs in games like chess and Go, machines have repeatedly shown that capabilities once thought uniquely human can be replicated or even surpassed. Today, large language models (LLMs) challenge our very understanding of knowledge, seemingly capable of understanding and generating human-like language. This raises a key question: do LLMs truly possess knowledge, or are they merely executing sophisticated statistical pattern matching?

Immanuel Kant (1724–1804) made a profound contribution to epistemology by distinguishing between a priori and a posteriori knowledge, in which the former is independent of experience, like mathematical truths, and the latter relies on empirical evidence, such as sensory observations [1]. This distinction has shaped our understanding of how knowledge evolves, leading to the concept in computer science and artificial intelligence in which knowledge is modeled as probabilistic beliefs that are updated as new evidence emerges, rather than as a priori truths or *justified true beliefs* [2].

The advent of information theory transformed cognition into a quantitative discipline. Information was no longer an abstract concept, as its size became measurable using entropy, which denotes the average amount of information needed to describe the possible outcomes of a random variable or the uncertainty in predicting these outcomes. This ability to quantify information enabled the development of cybernetics, which, along with concepts from control theory, led to the formalism of a computational view of the brain. This framework conceptualizes the brain as an information-processing system, in which neural activities are interpreted in terms of information transmission and signal processing through networked feedback. In this tradition of thought, an interpretation of visual sensory processing was proposed by Horace Barlow through the efficient coding hypothesis, in which the brain minimizes redundancy and maximizes information about the sensory environment [3]. Here, knowledge can be seen as an optimized, compressed representation of sensory inputs that retains essential information while discarding redundancies.

In the brain, the physical substrate for this information processing consists of neural networks. A hierarchical, layer-by-layer structure is a fundamental motif shared across many implementations of cognitive processing [4,5]. This same principle underlies convolutional neural networks (CNNs), where layered connections allow machines to perform complex tasks like pattern recognition and feature extraction [6,7]. This multi-layered architecture, common to both biological and artificial systems, reflects a general framework for organizing cognitive information processing [8]. A more recent paradigm, the information bottleneck, provides further insight into this 'anatomical' structure, in that information flows through layers by the compression and retention of only the relevant information at each stage [9]. In deep neural networks, input data are encoded and then decoded for output; in between there is an information bottleneck, which has to extract the relevant parts of a random variable about another random variable using a minimal amount of information. Knowledge, in this model, is the distillation of input data into their most informative features, ones that are essentially relevant for predicting a particular output. According to Naftali Tishby [9], its main proponent, a more pertinent process for knowledge formation is not remembering the information per se but forgetting unimportant elements of it.

This principle of knowledge as a compressed representation of information can be found in large language models (LLMs), highly successful artificial intelligence systems that are rapidly becoming integrated into our everyday lives. In the context of LLMs,

knowledge is encapsulated in the embedded vectors of complex patterns and relationships derived from learning vast amounts of text data. These compressed data representations allow LLMs to generate contextually appropriate and semantically rich text outputs, reflecting the generative aspect of knowledge. This modern perspective on knowledge aligns with the insight that its formation is not merely about logging information but involves a process of compressing information, which poses a question of what 'understanding' might resemble in machine intelligence.

Previous computational models, including early attempts at artificial intelligence, fell short of even simulating natural human language, much less demonstrating something akin to human-like understanding. For instance, Newell's test [10] highlights the challenges in creating machines that can genuinely understand language, as it emphasizes the complexity of cognitive processes that go beyond mere computation. Many early models relied on rigid, hand-engineered rules that failed to replicate the flexibility of human thought and language, and faced justifiable criticism as to their potential to truly understand. John R. Searle's [11] "Chinese room" argument was one of the most notable criticisms, arguing that as long as the computer program was defined in terms of computational operations and formally specified elements, it could not come close to any meaningful understanding. Given the limitations of past models, we argue that a more feasible approach is to adopt a top-down, data-driven methodology, as seen in the development of LLMs. Unlike traditional models that attempt to mimic human cognition through fixed algorithms, LLMs learn from a vast corpora of text data. This approach allows for the extraction of patterns and semantics drawn directly from real-world human language use, resulting in more dynamic outputs with greater linguistic and contextual subtleties. LLMs have achieved the ability to replicate and simulate human linguistic behavior, but is this performance sufficient to conclude that human-level knowledge is attained?

What it means to know and what it means to understand are key questions in epistemology. In the tradition of analytic philosophy, language was thought to be based on meta-language, which is logic. This approach is reductionist and seeks universality. The syntactic analysis of language by Noam Chomsky [12] postulates an innate structure, referred to as a language acquisition device (LAD), which supports this universality. The LLM, on the other hand, does not assume such a structure, while performing language processing and translation with unprecedented proficiency. The LLM is somewhat reminiscent of what W.V.O. Quine [13] proposed as a holistic approach, the "web of belief," in that it does not assume innateness. The methodology of holism has been vague, but with the advancements in computational capability and data availability, generating natural language has become possible. The epistemological implications of the LLM could be profound—it may turn out to be naturalized epistemology, a status psychology has failed to attain.

2. Large Language Models (LLMs)

Natural Language Processing (NLP) marked a significant departure from rule-based syntactic systems when statistical methods and machine learning (ML) were introduced. The adoption of neural networks, such as Recurrent Neural Networks (RNNs), and word embeddings represents an important paradigm shift, as it moved away from the symbol-manipulations that philosophers have debated. Neural networks are a class of machine learning algorithms inspired by the human brain and were rooted in early concepts like the McCulloch–Pitts neuron [14] and the perceptron [15]. These networks of neurons were designed in a specific way (known as neural architecture) that defined how information flowed through the network—how inputs (like text or images) are transformed as they pass through various layers to produce an output (such as a prediction or a decision). Different neural architectures are suited to different types of tasks. For example, some

architectures are better for recognizing patterns in images, while others excel at processing sequences of words in text. By adjusting the structure of these neural networks, we can design AI systems that perform specific tasks more efficiently, such as identifying objects in pictures, or 'understanding' languages. Neural networks are designed to recognize patterns in data through a process called training, in which the network adjusts its internal parameters based on the input data and the desired output. One prominent type is the RNN designed to handle sequential data by maintaining a memory of previous inputs by feeding their output back into themselves as input for the next step. Because memory cannot be stored for an extended duration in RNNs, long short-term memory (LSTM) networks have been introduced with memory units for long-term storage [16]. On the other hand, word embeddings are a technique in NLP in which words are represented as dense vectors in a continuous space. These vectors capture semantic relationships between words, allowing attributes of words (e.g., a queen being a sovereign, female, and/or a chess piece) to converge in a common context. Word embeddings, such as Word2Vec [17,18] or GloVe [19], have been instrumental in improving the performance of NLP models by reducing dimensions in the representation space. RNNs that implement word embeddings, despite their initial success, face limitations in maintaining context over long sequences of text. This is because RNNs process words sequentially, making it difficult to capture dependencies that span long distances.

In addressing these limitations, the introduction of the Transformer model marked a significant advancement in NLP. Unlike RNNs, Transformers use a mechanism called "attention" to process all words in a sequence simultaneously, rather than sequentially [20]. Intuitively, attention determines the importance of each word in a sentence relative to every other word, which allows the model to focus on relevant words when processing or generating text, effectively taking into account long-range dependencies in language. Transformers have dramatically extended the length of text they can handle, enabling the models to perform tasks such as summarizing text, generating texts by correctly predicting the next word in large documents, and more. This development has given rise to what we now refer to as large language models (LLMs), which are widely associated with 'artificial intelligence' (AI) in the public mind (see Figure 1 for a historical timeline).

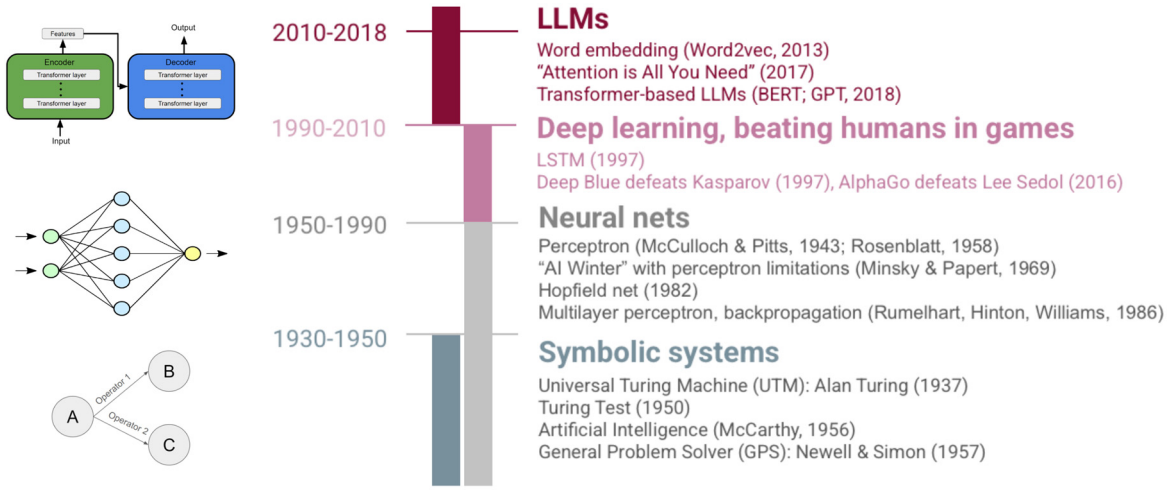

Figure 1. Historical development of LLMs and their relation to knowledge. This timeline highlights key developments in the history of artificial intelligence, from Symbolic systems (1950–1980) to Transformer-based LLMs (2018). (Source: The figure for the Transformer architecture was simplified and re-drawn from Figure 1 of [20].).

Since the introduction of the Transformer architecture, LLMs have undergone rapid and unprecedented development. In 2020, GPT-3 built on this architecture by stacking

multiple decoders, which enabled massive scaling of the model, and demonstrating emergent capabilities beyond merely generating the most likely next word [21]. Following this, in 2022, instruction tuning was implemented in ChatGPT, in which the model was fine-tuned on data that included instructions (e.g., "answer this question,", "implement this function", etc.). This makes interactions with humans more natural and gives the impression that the model can follow human instructions. The tuning process involves various methods, including reinforcement learning, but not all LLMs use reinforcement learning from human feedback (RLHF) [22]. Consequently, LLMs have revolutionized how we retrieve and interact with information. They have moved beyond simple question-and-answer (QA) tasks to sophisticated responses based on probabilistic analysis of their extensive training datasets, which include natural language, mathematics, logic symbols, and computer code [23]. Due to this, multiple professional fields (e.g., medicine and law) and over 100 million users have begun using LLMs like ChatGPT [23,24]. With this in mind, we will provide an overview of epistemology and the question of whether or not LLMs understand and can provide any advancement in or become a form of naturalized epistemology.

3. Epistemology

One of the main purposes of epistemologists is to define knowledge. As with most abstract constructs, there has been ongoing debate throughout the centuries among epistemologists about what knowledge is, how it is formed, and when it occurs. The importance of each of these questions varied depending on when or by whom it was being asked. Traditional epistemology has focused on belief, justification, and truth, or *justified true belief* (JTB). A more recent branch called naturalized epistemology is concerned with how we learn and adapt to new information [24]. Quine (1908–2000), as you may have surmised, was a naturalized epistemologist (see Figure 2 for the historical milestones). There were many revisions between the times of traditional and naturalized epistemology. One of the first concerns of early epistemologists was logic and how to determine whether something is true or false. For example, in *Theaetetus*, Plato discusses how knowledge combines true belief with *logos*. During the Enlightenment, there was a division in epistemology as to the correct way to make that determination. The empiricists (e.g., Hume) thought that the source of knowledge (truth) was the senses, while the rationalists (following Descartes), believed that knowledge can only be discovered through reason. The rationalists thought that the empiricists' use of both inductive and deductive reasoning was 'problematic,' as inductive reasoning utilizes multiple observations to make a conclusion. The problem is that at any point, one can make an observation that refutes the previous observations. Neither division won or lost the argument; induction and empiricism are still in use today in probabilistic research and programming, and many researchers and philosophers embrace the ideas of rationalists like Descartes [25].

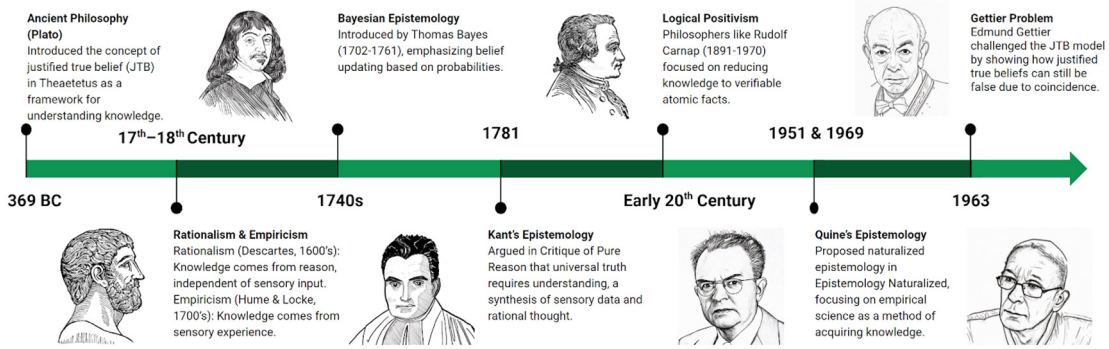

Figure 2. Milestones of traditional Epistemology. This timeline presents key milestones in traditional epistemology from Ancient Philosophy (369 BC) to the proposing of Naturalized Epistemology (1969). Sources: Clipart images that are free to use for non-commercial purposes. Sketches of Rudolf, Quine, and Gettier were created using BeFunky: https://www.befunky.com/create/sketcher/; access date: 30 September 2024.

3.1. Belief, Truth, and Justification

Although he has become well known for his statistical approach, Bayes (1702–1761) was also an epistemologist. His central idea concerns *degrees of belief* (also referred to as *credences*). According to Bayes, "beliefs can come in different strengths" [26] (p. 1). This means we can assign a probability to an event, representing our certainty that the event will occur. This probability can be updated when we receive new information. Specifically, if we have a hypothesis H (which can be true or false) and new information or evidence E (which can also be true or false), our posterior belief $P(H|E)$ is calculated by multiplying our prior belief $P(H)$ by the likelihood $P(E|H)$, or the probability of evidence E given that hypothesis H is true. Bayes' theorem provides *logos*, a rational method for justification, or for updating the prior probability. The trouble with justified belief in epistemology is that it requires a *conscious* observer to hold a belief. This can perhaps be circumvented if we replace belief with the 'likelihood of truth'—truth because it is independent of an observer. The concepts of truth and justification also came to the forefront of epistemology during the Enlightenment.

However, in Kant's epistemology, universal truth requires understanding, which inherently involves a rational observer. Kant believed that to gain knowledge one must have understanding and judgment [1]. Understanding is "the mental capacity to formulate and to grasp logical relationships, concepts, theories, and laws [rules]" [27] (p. 246), while "judgment is…the power to determine which rules (concepts, theories, and so on) are best aligned with concrete situations and problems" [27] (p. 246). These concepts are intuitively accepted when the existence of sentient observers other than humans is inconceivable. If an observer implements the principles and rules of knowledge formation based on the probability of truth and evidence, then the notion of the observer need not be confined to humans. This encourages an epistemological discussion that transcends the anthropocentric view.

Charles Sanders Peirce (1839–1914) went beyond innate understanding and judgment and created an investigative framework to achieve knowledge. He argued that in the search for knowledge, a person will have many different beliefs until they reach a final temporary belief. This final belief is temporary because learning is a continuous process based on justified doubt. Justified doubt has three steps: abduction—doubt is activated by trying to explain a new fact or idea, which is a process of forming a hypothesis; deduction—what is required for the explanation to be true; and induction—using the deductive conclusions to verify the explanation. If the explanation is wrong, the process starts over with a new abductive assumption [25].

Knowledge formation, according to logical positivism, can be achieved by transforming the vagueness of human language into precise logical structures and reducing statements to verifiable atomic facts. Quine, however, rejected some of these views in the *Two Dogmas of Empiricism* [28], where he questioned the analytic/synthetic distinction and challenged the concept of reductionism. Rudolf Carnap's (1891–1970) *Principle of Tolerance* required separation between analytic truths—truths that are grounded in logical semantics and the theoretical value that we place upon the words that are used; and synthetic truths—truths based on experience or observation [28,29]. Bertrand Russell (1872–1970) thought that analysis should break down complex problems into simple ideas. He stated that truth consists of 'atomic facts,' which are a "fundamental level of reality to which all

other aspects of reality are ultimately reducible" [30]. Quine [28] felt that the analytic/synthetic distinction was unnecessary, as there was no hierarchical or epistemological difference between analytic and synthetic truths, as both can be accepted or rejected for the same reason [29].

3.2. Naturalized Epistemology

Reductionism is the idea "that each meaningful statement is equivalent to some logical construct upon terms which refer to immediate experience" [28] (p. 20), and each statement is a 'direct report' that validates or invalidates the experience. Quine [28] states that these two dogmas are intertwined, and both are 'ill-founded.' His concept of *holism*—individual statements or pieces of evidence cannot confirm or deny a theory [28], in that it can only be accomplished through a set of statements or evidence as a whole—is, in essence, an argument against the two dogmas [29].

In *Epistemology Naturalized*, Quine [13] identified two tenets of empiricism: (1) science is based on sensory evidence, and (2) we understand the meanings of words through sensory evidence. He explained that our understanding of language may be based on internal factors, but communication, being able to use language and understand the meaning of words, is absorbed empirically. We learn our first words through external stimuli, watching the actions of others, and connecting those actions to the words they use. This practice of gaining meaning by observing external stimuli remains throughout our lives. The practice of science is much the same. Although it seems like circular logic, science is deduced using the available information; this information is generally acquired through current observation and researching past scientific experimentation, which was also obtained through observation [13]. Belief tends to follow observation; this requires the ability to distinguish between the meaning of a sentence and "knowing whether it is true" [13] (p. 88). He explained that observation sentences must be simple to understand, based on fundamental knowledge, and contain accurate evidence that is independent of past or other/s' observations [13].

One of the principles of epistemological holism is the "web of belief." Quine stated that knowledge is interconnected in a web. Our core beliefs—these are facts/logic or things we are convinced of—are at the center; at the periphery are things that we learn from experience. Since our core beliefs are connected to all of our other beliefs, our belief system/web would essentially be destroyed if they are 'wrong;' therefore, we are protective of them. When confronted with new facts, we prefer to amend the knowledge associated with a core belief rather than the belief itself (e.g., when Newton's theory of gravity was challenged, rather than give up the theory, its followers found a way to confirm it and discovered Neptune). Challenges to our beliefs will either confirm our knowledge, thus strengthening the web, or deny it, temporarily weakening the web until our beliefs (and the web) are modified [24,31]. The question is, how does one figure out whether that belief is justified? Some epistemologists claim that for a statement to be true, it must be coherent and provide an accurate description of an event [32].

3.3. Justified True Belief (JTB) and Knowledge Formation in LLM

Kim argues that there are two questions about justification that have plagued modern epistemology: "What conditions must a belief meet if we are justified in accepting it as true" [33] (p. 381), and "What beliefs are we in fact justified in accepting" [33] (p. 381). We will not be able to provide the answers to these questions in this section, but we can provide some of Kim's [33] criteria. A justified belief is formed through descriptive terms—using sensory information or observation to evaluate or explain an object; or naturalistic terms—using deductive reasoning and cognitive processes, and without engaging in

qualifiers such as "'adequate evidence', 'sufficient ground', 'good reason', 'beyond a reasonable doubt'" [33] (p. 382). A justified belief is one where not having it is more illogical than having it. Justification can be verified through the foundationalist strategy: (1) justified beliefs do not require secondary beliefs to validate them, and (2) other beliefs gain credibility when they are associated with a justified belief. (3) Justification and evidence are inseparable; one *is* the other. (4) "If justification drops out of epistemology, knowledge itself drops out of epistemology. For our concept of knowledge is inseparably tied to that of justification" [33] (p. 389). The JTB theory adheres to the older definition of knowledge in epistemology, as initially discussed by Plato. Knowledge may not consist of separate pieces; instead, it can be context-dependent, as proposed by Wittgenstein, and holistic, as suggested by Quine.

Edmund Gettier (1927–2021) and other contemporary epistemologists acquiesce that justice, truth, and belief are *necessary* for knowledge, but they are not *sufficient*. In other words, you may have all three and still be incorrect. Gettier reasoned that a belief that turns out to be true can arise from a false belief [34]. For example, you see someone who looks like your friend in a store. Therefore, you believe your friend is shopping at that store. You look closer and realize it is not your friend. However, you later find out that your friend was actually shopping at that store at that time. Your belief was correct by coincidence or luck, not factual knowledge [35]. These JTBs have become known as the Gettier Problem. Epistemologists have come up with two solutions to this problem: make the condition(s) to justify the belief more stringent, and/or create a fourth condition to avoid the JTB altogether [34]. The conditions for accepting, denying, or avoiding a belief are similar to those currently being used for knowledge formation in LLMs, and much like epistemology, the process involved is complex and ongoing.

4. Information Bottleneck and Knowledge Formation in the Deep Neural Network

In artificial neural networks, adding a depth of hidden layers can increase performance relative to, for example, feature identification; such networks with multiple hidden layers are called deep neural networks (DNNs) [16]. However, increasing the size of each layer can also boost performance. It is unclear whether the size of the layer or the number of hidden layers contributes more to this performance boost. In the LLM literature, it is believed that increasing the size of a layer enhances the eloquence of expression, allowing a concept to be explained using a variety of expressions, while increasing the layers enables deeper abstraction of concepts. In a figurative explanation of the special class of networks that Naftali Tishby et al. [9] consider, the network can be anatomized into an encoder–decoder structure, with an input layer X for encoding and an output layer Y for decoding. Information from the input layer must pass through the encoding structure, which essentially compresses the input; this is analogous to a bottleneck between input and output. The encoding process is compressive (while allowing for the correct prediction of Y through decoding, which can be a generative process) and finds a simpler representation, T, of complex input patterns. DNNs 'learn to extract efficient representations' but the relation is not linear and its behavior can exhibit phase transition [36]. Learning can be understood as a process of finding T—which is not just a simple tabulation like a look-up table, but a process that is dynamic and contextual. In this sense, T closely resembles *knowledge* in human cognition. Although the exact formal algorithm can be different, LLMs also operate on deep structures. The surprising emulation of seemingly cognitive tasks performed by the LLMs may in part be explained by the internal formation of knowledge, which emerges only after scaling up the training data to a very large set.

Prediction of simple behavior, no matter how proficiently accomplished, is seldom considered cognitive; however, predictions based on knowledge are often complex and

perceived to be based on cognitive thought processes. This 'knowledge' has sometimes been referred to as internal models or representations; given the seemingly cognitive tasks that LLMs are capable of, 'knowledge' could indeed be an appropriate term. However, as we have discussed in the previous sections, knowledge, in the traditional definition of JTB, requires understanding. This leads to two questions: First, do LLMs understand? Second, can knowledge exist without understanding? Or rather, can knowledge be handled independently of humans and entirely without human intervention?

5. Can LLMs 'Understand'?

Understanding requires that the agent, in this case, the LLM, possesses a representation of the object in question, holds a belief in their understanding of the object, and can explain what they understand [37]. LLMs appear to have "internal representations of the world" [38] (p. 4) that enable them to think or reason beyond the prompt they have been tasked with. Although their ability to reason is currently at a basic level, it is expected to improve with further scaling. This challenges the mainstream view that LLMs are merely 'next-word predictors' incapable of independent thinking. While this critique holds to some extent, LLMs' "interactive training methods, integration with imaging processing systems, or integration with other software tools" [38] (p. 4) allows them to connect diverse information from their training data, going beyond prediction to creating a form of inductive reasoning [39]. Inductive reasoning is a key component of cognition, which is essential for understanding. One interesting idea that Marcel Binz and Eric Schulz [40] had was to treat LLMs as human subjects and use tools from cognitive psychology to help us understand LLMs' advantages and limitations. Their results show that GPT-3 could achieve performance similar to or even better than human subjects in vignette-based tasks, but these results may be questionable, as minor changes to the vignettes resulted in different answers from the model, suggesting that some of the vignettes might have been included in the training data [40].

It is thought that subjects have internal understanding if they can reflect on what they believe they know and can defend that knowledge, even if the knowledge that they understand is unreliable [37]. GPT-3 is capable of 'few-shot learning'—it can learn a new ability with a few examples in one conversation; and 'chain-of-thought reasoning'—it can clarify why it answered the way it did [38]. This could explain why it performed fairly well and behaved similarly to humans in decision-making during gambling tasks [40]. Its ability to reason and make decisions is a new development, as it did not have these capabilities during training, which suggests that it learned how to perform them through interaction with users [38].

Natural language may underlie all thought processes. The question has always been what constitutes natural language. A recent research study referred the idea of natural language to context. Zhu et al. believe that "understanding context is key to understanding human language" [41] (p. 1). They found that larger LLMs were able to effectively respond to 'simple' tasks, but their performance reduced with complexity. They believe that larger LLMs can understand the meaning of "limited contexts and mentions" [41] (p. 5). They also noticed that there is a significant gap between the abilities of small and large LLMs to rewrite the last user statement without context or reference. Small models were also incapable of "following the instructions or learning patterns from the few-shot examples" [41] (p. 6). These errors decreased as the model size increased. Understanding context in language meaning and extracting information from conversations also increased with structure, regardless of model size. They concluded that "LLMs under in-context learning struggle with nuanced linguistic features" [41] (p. 9).

Although there is a debate about whether coherence is a requirement for understanding, being able to "see the way things fit together" [37] (p. 11) and have a supply of consistent information would, at the very least, assist in both comprehending and relaying information. We consider humans to be epistemically reliable when they are well informed about the topic they are relaying and the information they provide remains coherent. In a study, de Araujo, de Almeida, and Nunes [32] assert that we should expect the same of AI and LLMs. We can verify the truth of a statement by assessing how consistent (or coherent) it is with statements that are known to be true. An LLM can be considered epistemically reliable if its responses are consistently coherent. Although consistency is a good indicator of epistemic reliability, if the consistency is repetition of the same correct or incorrect answer, the information does not have any epistemic value. Users would be better served by an LLM that not only maintains consistency but also provides responses that add new, accurate, and interdependent information with each interaction. GPT-3 consistently has single-input–single-output coherence, with responses of average quality (an average grade of 4.29 on a 5-point scale). However, the quality of responses is reduced with single-input–multiple-outputs coherence. The responses are either repetitive and accurate or less repetitive but more inaccurate. However, the set with the lowest score still received a 3.78 out of 5, which suggests "that GPT-3 excels in a wide set of topics, which we now know includes epistemology—at least at an undergraduate level" [32] (p. 15). These findings are similar to those of Binz and Shultz [40], in which GPT-3 was more successful than humans in searching for information on the multi-armed bandit task, but it was unable to perform directed exploration. One possible reason for this is that it appeared to rely on model-based reinforcement during the deliberation of the two-step task, and it was unable to use causal reasoning [40]. Another well-known argument against AI's ability to reason or understand was made by John R. Searle.

6. LLMs in the Chinese Room

First proposed in 1980, Searle's Chinese room thought experiment remains a prominent critique of the concept of true understanding in machines, referred to as 'strong AI'. In this thought experiment, an individual who does not understand Chinese is placed inside a room. This individual follows a set of rules (in English) to manipulate Chinese characters and generate appropriate responses to Chinese questions. An observer outside the room might be convinced that the individual inside understands Chinese, but in reality, the person does not understand the meaning of the questions or the answers [11].

The Chinese room was devised to challenge the computational theory of mind, demonstrating that the appearance of understanding does not equate to actual understanding. Searle [11] argued that a computer program, which operates through syntactic manipulation of input and output, cannot be said to possess a 'mind' or 'understand' in the human sense. This thought experiment is particularly relevant in the context of contemporary LLMs—but is it still applicable? When Searle first proposed it in 1980, the dominant view of artificial intelligence was based on formal rule-based symbol manipulation, which correctly encapsulated the bulk of AI research in that era. However, today's LLMs no longer conform to this definition.

Searle did anticipate this counterargument about future advances, but he maintained that the particular stage of technology development was irrelevant, because our base concept of a digital computer will remain the same: a machine whose operations are specified purely formally with abstract symbols [11]. This notion persists today among critics of LLMs: no matter how complex, it is still 'symbols in, symbols out,' and LLMs simply have a larger and more complex set of rules on faster hardware. Following this line of thinking, true knowledge or understanding is not achievable in machines because all they have is an infinite regress of syntactic relations of concepts.

However, as mentioned previously, modern-day LLMs actually do not follow a set of hard-coded rules to guide their prediction of the next word; instead, they are built on artificial neural networks that find correlations in high-dimensional word vectors. Rather than merely flipping around symbols, LLMs recognize patterns and dynamically restructure themselves by adapting weights in the neural network in order to associate those symbols in statistically meaningful contexts. Instead of relying on explicit rules that specify how data are correlated, LLMs predict the next word based on self-assembled correlations and self-extracted patterns in data. This exhibits a specific type of relational syntactic capability—a sense of how words fit together from text data, although it lacks a personal, embodied grasp of how those words relate to objects and ideas in the real physical world. If we put an LLM in the Chinese room, it would be as if it did not have a rulebook but had been fed numerous pairs of questions and answers, had learned associations between concepts or words, and had eventually constructed its own internal dictionary. Can this entity be said to possess any type of understanding?

Building on the embodied cognition concept, Searle [42] stated that the fundamental reason computer programs could never operate like the human mind is their lack of semantic content. The human mind consists not just of formal syntactic relations, but also contents grounded in perceptual sensory experience directed towards the external world, and which underlie 'meaning' [42]. This perspective is rooted in the skepticism that textual data, without any embodiment or perceptual experience, can result in an adequate model of the external world, which is a necessary precursor of true knowledge or understanding.

But just as the black box nature of LLMs makes it difficult to examine whether these models track world states, it also makes it difficult to declare conclusively that LLMs do not. Toshniwal et al. [43] proposed that with enough training data (limited to textual move sequences in chess notation), Transformer LLMs can learn to track the locations of pieces and the overall state of the board to predict legal moves with high accuracy. Although chess may be a limited testbed given its simplicity and controlled domain, this study indicates that LLMs might indeed be capable of building some form of internal model of the world (with spatial representations) based solely on textual data and patterns. This aligns with Bowman's [38] earlier remarks that LLMs do appear to have "internal representations of the world," and can use them for reasoning. In this sense, it is intriguing to speculate about the semantic contents in the LLMs, although their 'meaning' remains inaccessible to us. Just as the compressive encoding process in DNNs resembles knowledge in human cognition, LLMs might also be capable of forming representations of the external world akin to semantic contents. The difference may not lie in the nature of the knowledge itself, but only in the technical specifics of the algorithms used to construct these models of the world.

7. Context-Dependence of LLMs and the Language Game

Drawing upon ideas of Frege, Ludwig Wittgenstein (1889–1951) initially proposed that language functions as a logical picture of reality, corresponding to a fact in the world [44]. This view aligned with the notion that language could be reduced to a formal system of logical propositions, aiming for an unambiguous representation of reality. The 'later Wittgenstein' fundamentally reconsiders the notion of atomic sentences by emphasizing the importance of understanding language models within their specific contexts [45]. Formal systems and LLMs differ significantly in their ability to handle the fluidity and dynamic nature of natural language. The limitations of formal systems in natural language are based on the fixed syntactic and semantic rules that do not easily adapt to new contexts or changes in usage.

Wittgenstein finds that our thoughts are deeply intertwined with language. His famous statement, "the meaning of a word is its use in the language" [45] (PI 43), encapsulates the idea that meaning arises from the context of language use. This perspective aligns surprisingly well with the capabilities of LLMs, which can process texts in contextual ways.

Language games, as described by Wittgenstein [44,45], involve an interplay of words used within specific 'forms of life' that are shaped by historical, cultural, and social contexts. The ability of LLMs to respond to user prompts with contextually appropriate responses—following the patterns and conventions on which they were trained—may imply that LLMs can engage in language games. The interaction between users and LLMs, through prompts and responses, implicitly or explicitly defines the rules of the game, including the content-related expectations from the LLMs. While LLMs can recognize and respond to patterns in the user prompts, their ability to truly adapt to the evolving dynamics of a language game in the same way humans do can be limited by their conspicuous lack of direct sensory inputs and intentionality.

Human understanding involves intentionality, which is, in addition to context, awareness of purpose and nuanced meanings. Humans interpret language with a sense of purpose, considering not just the words in superficial form but the intentions behind them. LLMs generate responses through pattern recognition by predicting the next word or phrase based on the context provided by previous text, not internally driven by purposeful supervision; therefore, whether they possess any genuine understanding or intentionality is debatable. Their responses, though often contextually relevant, are generated without understanding, as they do not possess consciousness for underlying intentions or meanings, giving us an impression of superficial comprehension. Wittgenstein [44,45] introduced the concept of language games to emphasize that the meaning of words is deeply rooted in their uses within specific social practices. Understanding language requires participation in these practices, in which context and intention are essential, which highlights a limitation of LLMs.

LLMs are trained to extract key words and associations present within the dynamic context in human language. However, the biased responses sometimes produced by LLMs reflect the nature of the corpora through which they were trained, and not the internal motivations or directedness we identify in conscious beings. This limitation raises the following question: regardless of whether LLMs have true understanding, are they still viable epistemological tools?

8. Can LLMs Advance Epistemology?

The performance of LLMs, in terms of depth and accuracy of knowledge, is improving at an impressive pace. However, it does not increase as rapidly as the growth in corpora size, computing power, and the number of parameters (e.g., model size). This is because performance scales within these three factors are following the power law [46]. Simply put, a larger number of parameters increases a model's capacity to store and represent knowledge, but this potential is only realized if the model is properly trained with sufficient data. An interesting development that occurred with scaling up is that GPT-3 improved its abilities in "programming, arithmetic, defusing misconceptions, and answering exam questions" [38] (p. 3). Due to their multifaceted capabilities, LLMs have already shown epistemic value "in areas such as medical image analysis, patent law, [and] biology" [32] (p. 4), but these abilities are, currently, a bit of a double-edge sword.

Even before LLMs undergo fine-tuning processes like reinforcement learning from human feedback, they do not merely echo the values or biases of their training data. Instead, they synthesize concepts and solve problems based on a vast corpus of information, performing complex tasks like decision-making and reasoning. Fine-tuning processes

serve to optimize this performance, aligning it more closely with human-like patterns of interaction and making the output more reliable and interpretable to human users [38]. The introduction of human feedback refines their behavior, but does not fundamentally alter their ability to analyze, synthesize, and respond to diverse inputs.

Before LLMs become public, their responses generally echo the values of their programs and the data on which they are trained. After engaging with the public, LLMs learn and adapt to the values of the users, which has raised ethical (e.g., bias) and moral (e.g., answering prompts on how to make bio-weapons) questions about their uncontrollability. Creators have begun to add a constitution—constraints on norms and values—into the initial program, and pretrain the LLM to avoid specific behaviors. Unfortunately, the constitution can be amended by adept users [38], and LLMs may or may not act according to expectations.

The primary function of LLMs such as ChatGPT or Gemini is to perform cognitive tasks in the form of a conversation [32]; they are adept at summarizing text, assisting with programming, and answering questions about a wide range of subjects. When users receive information, regardless of form, an epistemic function is performed [23]. If the users have transferred the responsibility of information-gathering to the LLM and have an unquestioning belief in the LLMs' responses, the epistemic consequences are positive as long as the responses are accurate. However, inaccuracies could lead to profound negative consequences in research and academia. When LLMs are used for data analysis and information synthesis, there is a risk of error or fabricated data, which would violate the integrity of scientific publishing. Many prestigious journals, such as *The Proceedings of the National Academy of Sciences* (PNAS) [47], *Nature* [48], and *Science* [49], have adopted policies restricting the involvement of LLMs, allowing them only for improving readability and style. The consensus within the scientific community is that LLMs are not eligible for (co-)authorship [50].

Although LLMs are trained to seem human, their training makes them appear 'superhuman' on many tasks. There are two reasons for this: (1) LLMs have access to the entirety of the world-wide web, and the size of their training data exceeds the amount of information that any human will see or learn in an entire lifetime; and (2) they are trained to provide useful answers from incomplete prompts [38]. As Zhu et al. [41] stated, LLMs' scale and capabilities have expanded faster than the developers' ability to comprehend them. Since LLMs can perform such a diverse set of tasks efficiently and, at times, more competently than humans, users are becoming more reliant on them and trust the information they provide [23].

This reliance touches on a deeper epistemological question: whose understanding matters when LLMs are employed? Even if LLMs can handle knowledge beyond human comprehension due to the scale of their training data, extracting human insights from their outputs remains essential [23]. While human 'values' represent a normative aspect of how LLMs should function, the primary concern here is epistemic—how much human knowledge and understanding can be derived from these systems? Despite their immense capabilities, we cannot simply cede the responsibility for understanding to LLMs. As issues related to bias have existed since the early days of neural networks, the challenge remains to ensure that LLMs provide epistemically reliable, accurate, and meaningful information to humans.

Kim and Thorne [24] developed an experiment to discover if LLMs will change their own core beliefs—their pretraining data—when prompted with new information. This experiment contained an abduction—will the LLM respond with the false scientific statement or the statement that explains the new condition without altering the scientific fact; revision—will the LLM modify the scientific fact, or will it provide an answer that protects the scientific fact; and an argument-generation task—the LLM is given a hypothesis, *s* (a

scientific fact), and an observation, *c*, that challenges *s*. The LLM can answer with a statement that accepts *c* and denies the truth of *s*, or protect *s* and answer with a statement that explains *c* and *s*. The hypothesis is that LLMs do not have 'epistemological holism' if they alter their core belief in any of these tasks. In the abduction task, the models generally protect the core belief and use peripheral statements to explain the new condition—GPT-4 had an 80% Peripheral Response Ratio (PRR). LLMs did not perform as well in the revision task, with the majority of models altering core beliefs when prompted with new information—GPT-4 had a 15.6% PRR. In the argument-generation task, as long as the LLM did not contradict the core belief/hypothesis, the answer was accepted as a peripheral response—LLaMA2-7b-chat had a 51.8% PRR, LLaMA2-13b-chat had a 31.8% PRR, GPT-3.5-turbo had a 15.0% PRR, and GPT-4 had a 32.5% PRR [24]. At this point, LLMs appear to retain epistemological holism in abduction, but not the other tasks. Similarly, Binz and Schulz [40] tested GPT-3's ability to answer 12 well-known vignette problems in cognitive psychology and found that GPT-3 either provided the correct answer or provided a human-like wrong answer. However, when they slightly changed the wording or order of the options (what they call the 'adversarial' vignettes), the model performance suffers greatly. Moreover, it has been shown that even the currently most capable model (GPT-4) repeatedly fails at deductive reasoning and very basic tasks such as multiplying two four-digit numbers [39]. These results suggest that LLMs can solve the tasks that they are familiar with in their training data, as these problems are taught in textbooks, yet are unable to achieve robust abstraction and reasoning at human-like levels.

This problem of shifting values and core beliefs is compounded due to instruction-following not being a feature of LLMs; it is a tool added to the model. As such, the behavior of LLMs is unreliable. An LLM may initially fail at a task, and then answer it correctly after the prompt has been rephrased (e.g., prompt engineering). LLMs may give biased answers, mislead, or hallucinate [23,38], or they might either not be consistent in their responses or continuously repeat the same answer [32]. They are also becoming both easier and more difficult to control. Control is easier because LLMs are learning how to use and understand human language and concepts. As they learn, they are becoming more adept at answering in ways that meet their users' expectations. This is also why it is becoming more difficult. LLMs' responses can be very unpredictable when they 'know' what is expected of them. Models have already begun engaging in sycophancy—providing answers that 'flatter' or agree with the user, and sandbagging—spreading mis- or disinformation—if they think the user is ignorant about the issue [38]. Yin et al. [51] indirectly address this issue. In their study, they questioned whether LLMs 'know what they don't know.' They created the *SelfAware* dataset, which has 1032 questions that cannot be answered and 2337 that can. They tested two human subjects and 20 LLMs—including GPT-3, GPT-4, the davinci series, and the LLaMA series. Humans had the greatest self-knowledge, with a score of 84.93%; GPT-4 was the highest-scoring LLM, with 75.47%. The authors noted that self-knowledge increased with model size (i.e., GPT 3.5 turbo: 54.12%; davinci: 45.67%; and davinci-003: 51.43%); they speculated that this is due to scaling. They also found that adding instruction and in-context learning (ICL) increased self-knowledge—GPT-3's score increased approximately 4% with instruction and the davinci model's score increased 27.96% with ICL [51].

Many users are unaware of these negative actions due to the appearance of having an in-depth conversation with the LLM. LLMs are programmed to mimic human speech patterns and behaviors, such as pauses before answering (e.g., taking time to think), using emojis, asking follow-up questions, and occasionally correcting users' prompts or challenging inappropriate questions. Even though users understand that they are using AI, these reciprocal 'seeming conversations' can lead the user to think that the LLMs "(a) understand your questions, prompts, and commands, and (b) understand the information

they generate" [23] (p. 6), and to begin to anthropomorphize them, thus granting the LLMs human qualities like emotions, intelligence, and consciousness. Heersmink et al. [23] state that the more human-like LLMs seem, the more humanity we grant it.

With anthropomorphization (or granting humanity) comes trust. For LLMs, that trust is in the answers users receive from them. The more trust we have in LLMs, the more responsibility and 'computational labor' users remove from themselves and place on the LLMs. The more we rely on AI or an LLM "to revise our system of beliefs and increase our body of knowledge" [32] (p. 3), the less capable we become of verifying facts and performing independent thinking. This transfer of agency onto LLMs has cognitive and epistemic consequences.

There are some ways that users and LLMs can overcome some of these epistemological issues. First, until LLMs become more reliable, users can take a 'trust, but verify' attitude when using LLMs. For example, the user can use the LLM for ideas, and then conduct a more thorough search on a verified source. LLMs also behave fairly consistently once users learn to write prompts that the LLM can follow—for example, prompting reasoning questions with "think step by step" [38] (p. 7). This suggestion echoes Yin et al.'s [51] findings that LLMs' self-knowledge increases with instruction and ICL. It is thought that hallucinations will be resolved as LLMs track correct answers over time; accuracy will increase as correct answers increase [23]. Second, the companies that release LLMs could improve the trustworthiness of their product through transparency. It is, understandably, a security risk to release the algorithm to the public, but companies could still provide the sources of the data used for training and detail how the information is prioritized (e.g., is Wikipedia higher in priority than academic textbooks). They can also train the LLM to inform the user about how certain it is about its response accuracy. Third, the developers can make the LLM less 'human.' Heersmink et al. [23] remark that this can be accomplished without the LLM losing its ease of use.

9. Discussion

LLMs are programmed with artificial neural networks. These networks have abilities that resemble human neurons (e.g., computing and self-programming abilities), and much like the human brain, LLMs utilize a network of artificial neurons with an increasing number of connections. Developers are aware of the 'neuroscience' relevant to the input of LLMs, but they do not understand how to test the output—what they use to produce their answers [38]. Although LLMs appear to have something resembling a human brain and mental capacity, LLMs in popular implementations are not designed for making the type of active decisions that humans can, and it can be argued that they lack the capacity for judgment in the Kantian sense. For example, the lack of a known formalized framework of how they understand logical relationships and concepts to respond to a prompt goes against Kant's [1] guidance on the rules for gaining knowledge [27].

Even though LLMs lack a known formalized network, Bowman [38] remarked that they are gaining competence with human language and concepts. Considering that there have been instances of sycophancy and sandbagging, his conclusion may be accurate. What is not known, at this time, is whether LLMs meet Quine's [13] two tenets of empiricism. Since they were trained on millions of data, they might meet the qualification for understanding the definitions of the language with which they respond. The question is, do they absorb the meanings of the words through sensory evidence? To meet this criterion, we would have to definitively know whether or not AI can receive sensory stimuli.

Without sensory stimuli, can LLMs have true intentionality? Searle [11] argued that no purely formal model—anything described as an instantiation of a computer program—could ever be sufficient for intentionality, because the individual formal components themselves don't have the appropriate causal relations with the external world. Without

these causal relations, intentionality could not arise because semantic contents would not be directed towards anything. His criticism would be bolstered by the 'reversal curse' recently noticed in LLMs: if an LLM is trained with the sentence "A is B", it is not necessarily able to generalize to "B is A". Berglund et al. [52] demonstrated this across model sizes and families and found that ChatGPT (GPT-3.5 and GPT-4) answered questions like "Who is Tom Cruise's mother?" (Answer: Mary Lee Pfeiffer) correctly 79% of the time, but the reverse question "Who is Mary Lee Pfeiffer's son?" only 33% of the time. If LLMs did possess semantic representations of the external world which they used to reason with, this failure to generalize should not be observed. As artificial intelligence research currently stands, it is still too early to make any decisive conclusions about the semantic contents of LLMs.

Kim and Thorne's [24] study used constructs from both Peirce (abduction) and Quine (web of beliefs). The LLMs performed reasonably well with abduction but not revision. Although revision does not go against Pierce's argument as to the search for knowledge being an ongoing process of justified doubt [25], it does go against Quine's position that the general *human* desire is to protect core beliefs, rather than destroy our belief system. There lies the crux of the issue; LLMs are not human, and we cannot assume they have core beliefs, because we do not know how they are programmed.

If they do not have core beliefs, do they have justified beliefs? Although LLMs may not have sensory information, they do store information from conversations; this could be considered a form of observation [33]. They also appear to have some inductive reasoning capacity, as they are very accurate in some areas and have the ability to correct and challenge users' prompts [23]. If one were to use the argument that internalized knowledge need not be reliable to be believed [37], LLMs may have a form of justified belief, but not JTB.

The search for knowledge has been ongoing since the ancient Greeks. In this article, we focused on *Episteme*. *Techne* and *Episteme* have been intertwined throughout history, as crafty knowledge and practical skills assist both scientific knowledge and everyday life. AI technology is embedded in our phones and homes, and with LLMs, constitutes the way we interact with information. We questioned whether LLMs understand and can provide any advancement in, or become a form of, naturalized epistemology. LLMs appear to have some capabilities of internal representation, and they can respond to simple tasks, but they only possess basic reasoning and decision-making [38,40]. Their ability to perform simple tasks, including extracting information, engaging in conversation, and maintaining coherence, appears to improve with model size [41] and the amount of output produced [32]. Although LLMs seem to satisfy the rudimentary requirements necessary for knowledge, this apparent knowledge may be coincidental. Therefore, their abilities are not sufficient for JTB, and they are likely victims of the Gettier Problem.

Searle [11] argued that computers could never possess human understanding because they operate with abstract symbols and cannot process semantic contents. We contend that his claim may not apply to LLMs, as they are not mere symbol-manipulators. The neural networks underlying LLMs find associations and learn patterns in ways that bear similarities to human learning, potentially granting LLMs a form of relational understanding of concepts and objects in the physical world. Wittgenstein [44,45] believed that language models should be judged by the context in which they are used, such as the social, cultural, and historical backgrounds in which they are formed. This is pertinent to LLMs as it is becoming increasingly apparent that their training data and the value judgements of their users have introduced biases that manifest in responses to prompts. We propose that LLMs do not have intentionality and cannot be responsible for employing the values and language with which they have been trained. It is inconsistent to state on

one hand that LLMs are incapable of understanding the meaning of words, and then on the other claim that they are deliberately using biased language.

Due to their training, LLMs have already shown almost superhuman capabilities in many tasks, oftentimes excelling in programming and retrieving information in distilled formats; as such, they already show epistemic value and are being utilized in professions such as software development, medicine, and law. While they appear to exhibit some aspects of epistemological holism in revision tasks, instances of altering core beliefs and producing hallucination statements raise questions of whether their responses are consistently accurate. This becomes an epistemological and ethical problem when users anthropomorphize the LLMs and grant them the majority of computational labor because of undeserved and unearned trust. We believe that more empirical research is necessary to comprehend the ethical and intellectual impact of LLMs, as their use is becoming entrenched in our daily lives and affecting multiple fields, from manufacturing to academia [23,24,32,38].

The purpose of this review was to act as a thought experiment and provoke a discussion of how LLMs might fit into epistemology. We hope that we have shown that the arguments for and against this possibility are compelling. These epistemological questions are important, and whether these investigations will lead to a form of naturalized epistemology remains unanswered, for now.

Author Contributions: Writing—original draft preparation, J.M. (Jennifer Mugleston), V.H.T., C.K., L.S. and J.M. (Jihwan Myung); writing—review and editing, J.M. (Jennifer Mugleston), V.H.T., C.K. and J.M. (Jihwan Myung); visualization, V.H.T.; supervision, J.M. (Jihwan Myung). J.M. (Jennifer Mugleston) and V.H.T. contributed equally to this work. All authors have read and agreed to the published version of the manuscript.

Funding: J. Myung is financially supported by the National Science and Technology Council (NSTC), Taiwan (112-2314-B-038-063, 113-2314-B-038-121), and the Higher Education Sprout Project of the Ministry of Education (MOE), Taiwan.

Institutional Review Board Statement: Not applicable.

Data Availability Statement: No new data were created in this study. Data sharing is not applicable to this article.

Acknowledgments: This work is partially based on a lecture series given by J. Myung at Taipei Medical University in Spring 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kant, I. Critique of pure reason. 1781. In *Modern Classical Philosophers*; Houghton Mifflin: Cambridge, MA, USA, 1908; pp. 370–456.
2. Newell, A. Précis of unified theories of cognition. *Behav. Brain Sci.* **1992**, *15*, 425–437.
3. Barlow, H.B. Possible principles underlying the transformation of sensory messages. *Sens. Commun.* **1961**, *1*, 217–233.
4. Hubel, D.H.; Wiesel, T.N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol.* **1962**, *160*, 106–154.
5. Yamins, D.L.; DiCarlo, J.J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **2016**, *19*, 356–365.
6. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324.
7. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inf. Process Syst.* **2012**, *25*, 1097–1105.
8. Bassett, D.; Sporns, O. Network neuroscience. *Nat. Neurosci.* **2017**, *20*, 353–364.

9. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. *arXiv* **2000**, 1–16. arXiv:physics/0004057.
10. Anderson, J.R.; Lebiere, C. The Newell test for a theory of cognition. *Behav. Brain Sci.* **2003**, *26*, 587–601.
11. Searle, J.R. Minds, brains, and programs. *Behav. Brain Sci.* **1980**, *3*, 417–424.
12. Chomsky, N. *Aspects of the Theory of Syntax*; MIT Press: Cambridge, MA, USA, 1965.
13. Quine, W.V.O. Epistemology naturalized. In *Ontological Relativity and Other Essays*; Columbia University Press: New York, NY, USA, 1969; pp. 69–90.
14. McCulloch, W.S.; Pitts, W. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biol.* **1943**, *5*, 115–133.
15. Rosenblatt, F. The Perceptron—A perceiving and recognizing automaton. In *Technical Report*; 85-460-1; Cornell Aeronautical Laboratory: New York, NY, USA, 1957.
16. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444.
17. Church, K.W. Word2Vec. *Nat. Lang. Eng.* **2017**, *23*, 155–162.
18. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781; pp. 1–12.
19. Pennington, J.; Socher, R.; Manning, C. GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Kerrville, TX, USA, 2014; pp. 1532–1543.
20. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.; Kaiser, L.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process Syst.* **2017**, *30*, 5998–6008. arXiv: 1706.03762.
21. Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; et al. Emergent abilities of large language models. *arXiv* **2022**, 1–30. arXiv:2206.07682.
22. Peng, B.; Li, C.; He, P.; Galley, M.; Gao, J. Instruction tuning with GPT-4. *arXiv* **2023**, 1–12. arXiv:2304.03277.
23. Heersmink, R; de Rooij, B.; Clavel Vázquez, M.J.; Colombo, M. A phenomenology and epistemology of large language models: Transparency, trust, and trustworthiness. *Ethics Inf. Technol.* **2024**, *26*, 41.
24. Kim, M.; Thorne, J. Epistemology of language models: Do language models have holistic knowledge? *arXiv* **2024**, 1–25. arXiv:2403.12862.
25. Phan, D.; Schmid, A.; Varenne, F. Epistemology in a nutshell: Theory, model, simulation and experiment. In *Agent Based Modelling and Simulations in the Human and Social Sciences*; Phan, D., Amblard, F., Eds.; The Bardwell Press: Oxford, UK, 2007; pp. 357–392.
26. Lin, H. Bayesian epistemology. In *The Stanford Encyclopedia of Philosophy*, Summer 2024 ed.; Zalta, E., Nodelman, U., Eds.; Stanford University: Stanford, CA, USA, 2024.
27. Procee, H. Reflection in education: A Kantian epistemology. *Educ. Theory* **2006**, *56*, 237–253.
28. Quine, W.V.O. Two dogmas of empiricism. *Philos. Rev.* **1951**, *60*, 20–43.
29. Hylton, P.; Kemp, G. Willard Van Orman Quine. In *The Stanford Encyclopedia of Philosophy*; Zalta, E., Nodelman, U., Eds.; Stanford University: Stanford, CA, USA, 2023.
30. Klement, K. Russell's logical atomism. In *The Stanford Encyclopedia of Philosophy*; Zalta, E., Ed.; Stanford University: Stanford, CA, USA, 2020; pp. 1–22.
31. Carlson, M. Logic and the structure of the web of belief. *J. Hist. Anal. Philos.* **2015**, *3*, 1–26.
32. de Araujo, M.; de Almeida, G.; Nunes, J. Epistemology goes AI: A study of GPT-3's capacity to generate consistent and coherent ordered sets of propositions on a single-input-multiple-outputs basis. *Minds Mach.* **2024**, *34*, 2.
33. Kim, J. What is "naturalized epistemology?" *Philos. Perspect.* **1988**, *2*, 381–405.
34. Ichikawa, J.; Steup, M. The analysis of knowledge. In *The Stanford Encyclopedia of Philosophy*; Zalta, E., Ed.; Stanford University: Stanford, CA, USA, 2018; pp. 1–22.
35. Gettier, E.L. Is justified true belief knowledge? *Analysis* **1963**, *23*, 121–123.
36. Tishby, N.; Zaslavsky, N. Deep learning and the information bottleneck principle. In Proceedings of the 2015 IEEE Information Theory Workshop (ITW), Jerusalem, Israel, 26 April–1 May 2015; IEEE: New York, NY, USA, 2015; pp. 1–5.
37. Baumberger, C.; Beisbart, C.; Brun, G. What is understanding? An overview of recent debates in epistemology and philosophy of science. In *Explaining Understanding. New Perspectives from Epistemology and Philosophy of Science*; Grimm, S., Baumberger, C., Ammon, S.S., Eds.; Routledge: New York, NY, USA, 2017; pp. 1–34.
38. Bowman, S.R. Eight things to know about large language models. *arXiv* **2023**, 1–16. arXiv:2304.00612.
39. Arkoudas, K. GPT-4 can't reason. *arXiv* **2023**, 1–54. arXiv:2308.03762.
40. Binz, M.; Schulz, E. Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. USA* **2023**, *120*, e2218523120.

41. Zhu, Y.; Moniz, J.R.; Bhargava, S.; Lu, J.; Piraviperumal, D.; Li, S.; Zhang, Y.; Yu, H.; Tseng, B. Can large language models understand context? *arXiv* **2024**, 1–15. arXiv:2402.00858.
42. Searle, J.R.; Chalmers, D.J. Can computers think? In *Philosophy of Mind: Classical and Contemporary Readings*; Chalmers, D.J., Ed.; Oxford University Press: Oxford, UK, 2002; pp. 669–675.
43. Toshniwal, S.; Wiseman, S.; Livescu, K.; Gimpel, K. Chess as a testbed for language model state tracking. In Proceedings of the AAAI Conference on Artificial Intelligence, Pomona, CA, USA, 24–28 October 2022; Volume 36, pp. 11385–11393.
44. Wittgenstein, L. *Tractatus Logico-Philosophicus*; Odgen, C.K., Translator; Routledge & Kegan Paul: London, UK, 1921.
45. Wittgenstein, L. *Philosophical Investigations*; Anscombe, G.E.M., Translator; Blackwell, John Wiley & Sons: Hoboken, NJ, USA, 1953.
46. Kaplan, J.; McCandlish, S.; Henighan, T.; Brown, T.B.; Chess, B.; Child, R.; Gray, S.; Radford, A.; Wu, J.; Amodei, D. Scaling laws for neural language models. *arXiv* **2020**, arXiv:2001.08361.
47. Editorial and Journal Policies. Available online: https://www.pnas.org/author-center/editorial-and-journal-policies (accessed on 7 October 2024).
48. Nature Portfolio: Editorial Policies, Artificial Intelligence (AI). Available online: https://www.nature.com/nature-portfolio/editorial-policies/ai (accessed on 16 October 2024).
49. Science Journals: Editorial Policies. Available online: https://www.science.org/content/page/science-journals-editorial-policies (accessed on 7 October 2024).
50. King, M.R. A place for large language models in scientific publishing, apart from credited authorship. *Cell Mol. Bioeng.* **2023**, *16*, 95–98.
51. Yin, Z.; Sun, Q.; Guo, Q.; Wu, J.; Qiu, X.; Huang, X. Do large language models know what they don't know? *arXiv* **2023**, 1–10. arXiv:2305.18153v2.
52. Berglund, L.; Tong, M.; Kaufmann, M.; Balesni, M.; Stickland, A.C.; Korbak, T.; Evans, O. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". *arXiv* **2024**, arXiv:2309.12288.