On eavesdropping octopuses and stochastic parrots: what do they know?

Henrique Gomes[1] and Vasudev Shyam[2]

[1]Oriel College, University of Oxford, UK
[2]Zyphra, Palo Alto, CA, USA

October 15, 2024

Abstract

The extant literature on AI (and popular culture more generally) has a few popular slogans that seek to dismiss the cognitive capacities of current large-language models (LLMs). Here, from a conceptual standpoint, we assess whether two such slogans have any teeth. The first such slogan is that "LLMs can only predict next-tokens". The second is that "AIs are stochastic parrots". We will briefly explain these two slogans, and argue that, in plausible construals, they do not imply fundamental limitations to cognition and semantic grounding (which of course does not imply anything positive about current AI's cognitive capacities). The difference between our approach and that of the burgeoning literature reaching a similar conclusion is that we base our arguments on the idea of 'knowledge-first epistemology'.

1 Introduction

As artificial intelligence (AI) systems, particularly large language models (LLMs), have advanced in recent years, they have sparked both excitement

and skepticism regarding their cognitive capacities. In terms of practical use, an LLM was until recently exclusive to a handful of aficcionados and researchers. In a matter of a couple of years, it has reached the masses, being used for administration of complex organizations, for planning different types of activities, for writing code and non-fiction, etc. Nonetheless, many take the tasks being automated by LLM's as merely automatable drudgery. These critics take current LLM's to suffer from fundamental, structural constraints in their cognitive capacities.

Two prominent slogans often used to dismiss the capabilities of LLMs are: "LLMs can only predict next-tokens" and "AIs are stochastic parrots." These slogans suggest that LLMs lack genuine understanding or intelligence, functioning merely as sophisticated statistical machines without true semantic comprehension.

The first accusation—that LLMs can only predict next-tokens—implies that non trivial cognitive tasks, such as 'understanding' language, or having complex world-models, lies beyond what LLM's can achieve. For instance, Melanie Mitchell (2019,0) argues that AI systems "do not capture the rich meanings that humans bring to bear in perception, language, and reasoning", and she calls this lack of understanding a 'Meaning Barrier".

The second, related accusation—that AIs are stochastic parrots—was notably articulated by Bender et al. (2021) and suggests that LLMs merely mimic language without any understanding, akin to parrots generating or repeating phrases without grasping their meaning, and using a stochastic process to determine the content of their output. This notion aligns with the concept of *semantic zombies*: entities that exhibit behavior indistinguishable from that of sentient beings but who lack any referential mechanism (cf. (Lyre, 2024) and references therein). The usual criticism here is phrased in terms of *the symbol grounding problem* for meaning.

Mollo and Millière (2023) summarise the core criticism as follows:

> if AI systems are merely designed to process linguistic inputs, how can their outputs be grounded in a world with which they have no direct interaction? How can those outputs possess any

meaning beyond the interpretations that we, as intelligent beings embedded in the world, may project onto them?

They distinguish between four kinds of grounding problem: referential, sensorimotor, relational, communicative, and epistemic grounding, and go on to argue that, once one takes into account human reinforcement feedback learning, all of them are overcome by current LLMs.

In this paper, we also aim to deconstruct these critical slogans. In this sense, we align ourselves with the more positive views on semantic grounding of (Chalmers, 2023; Mollo and Millière, 2023; Pavlick, 2023; Piantadosi and Hill, 2022; Søgaard, 2022,0).

But our reply to these slogans is different from previous ones in important ways. We first shift focus from semantic grounding to *knowledge*, which we take in the standard, Aristotelian sense of justified true beliefs. But here we will argue that knowledge about the world does not need *direct* causal contact with the primary objects of reference. As argued by Bird (2023), knowledge is, in a sense, 'Markowian': previous knowledge can serve just as well as evidence for further knowledge as ostentation, sensorimotor experience, or any sort of direct acquaintance. And human languages and large amounts of text data mirror the world in multiple ways and therefore contain knowledge about the world. A more detailed prospectus is as follows.

1.1 Prospectus

We will begin in Section 2 by demystifying next-token prediction, by explaining it as a kind of language parsing. Specifically, next-token is understood as a neural sequence model trained to predict, given a context the next part of a sentence, most commonly in units of syllables or even letters—the token. We argue that next-token prediction serves as a method for training models to demonstrate language membership by parsing derivatives of the training corpus. It is not so much the famous transformer arquitecture but the next token prediction that actually captures non-trivial structure of our body of texts. So next-tokens can be seen simply as the method to acquire language, and do not imply anything about its use. To end this Section, we draw par-

allels between this procedure and how humans learn language, noting that both involve quickly determining whether words or sentences belong to a language. This Section will include more technical details then the arguments of the following Section 3, since those have already been given in the literature (cf. (Lyre, 2024; Mollo and Millière, 2023; Pavlick, 2023; Søgaard, 2022,0)) and we will merely summarise.

In Section 3, we address a common justification for the accusation of "stochastic parroting", which we take to be essentially the 'symbolic grounding' claim, that AIs and LLMs lack direct acquaintance with the world, or sensorimotor grounding and are thus condemned to be "semantic zombies." Our main argument here draws from modern philosophy's dismissal of the value of direct acquaintance. We have no 'immaculate', 'direct' access to the world either: knowledge builds up from intricate 'coordination problems' at the roots of languages towards more complicated theories about the world, that need not mention these roots. Particular to our dismissal is Bird (2023)'s argument that knowledge—justified true beliefs—constitutes evidence, and that evidence is knowledge used in successful inferences towards more knowledge. And there is plenty of evidence that AI more broadly, but even LLM's, have indeed acquired knowledge from the data that we have supplied it.

2 Next-Token Prediction as Language Parsing

Tokens are words or parts of words that figure with some relative autonomy in the texts used to train LLMs. And LLMs are primarily trained using the objective of next-token prediction. Given a sequence of tokens (words, subwords, or characters), the model predicts the probability distribution over the possible next tokens. This training objective enables the model to learn the statistical patterns and structures present in the language data it has been trained on.

Mathematically, given a sequence of tokens $(w_1, w_2, \ldots, w_{n-1})$, the model aims to estimate the probability $P(w_n|w_1, w_2, \ldots, w_{n-1})$. By maximizing the

likelihood of the correct next token over large corpora, the model learns to generate coherent and contextually appropriate text.

2.1 Language Parsing by Derivatives of the Training Corpus

Next-token prediction can be viewed as a form of language parsing, where the model learns to recognize and generate sequences that are grammatical and semantically coherent within the language. By predicting the next token, the model effectively demonstrates membership of a sequence within the language, as it must understand the syntactic and semantic constraints that govern token order. And, the *process* by which LLM's acquire meaning, unlike other questions regarding meaning and indeed unlike the neural process by which humans acquire language, is very well understood. We now give a quick summary.

2.1.1 Language membership through derivatives

If we take a language L to be a set of strings formed from an alphabet A, i.e.

$$s \in L \Rightarrow s = w_1 w_2 \cdots w_n, w_i \in A \ \forall i. \tag{1}$$

Brzozowski defined the derivative of a language with respect to an element of the alphabet w_k (which can be characters or tokens depending on what alphabet we choose to form strings with) as the set of all w_k-suffixes in the language. Formally, we define the ∂_{w_k} operator as:

$$\forall s \in \partial_{w_k} L, \ \text{concat}(w_k, s) \in L. \tag{2}$$

Here, concat denotes the string concatenation operator. It's action on two strings $s_1 = w_1 w_2 .. w_k$ and $s_2 = w_{k+1} .. w_{k+n}$ reads:

$$\text{concat}(s1, s2) = w_1 w_2 \cdots w_k w_{k+1} w_{k+2} \cdots w_{k+n}. \tag{3}$$

Parsing with derivatives is the process by which the membership of a string s to a language L can be proved by taking subsequent derivatives with respect

to each alphabet element of the string and seeing if we are left with the empty *set*, , in which case the string does not belong to the language or if we end up with a language containing the empty *string*, ϵ, meaning the original undifferentiated string belongs to the language. The emtpy string is the string with no characters, which is the result of taking derivatives with respect to all characters/tokens in a string that belongs to the language. Any derivative of the empty string results in the empty set. The empty set is a set with no inhabitants. Formally

$$s = w_1 \cdots w_n \in L \text{ iff } \epsilon \in \partial_{w_n} \cdots \partial_{w_2} \partial_{w_1} L, \tag{4}$$

The property that the empty string belongs to a language is known as *nullability*. Note that we can choose instead to read strings right to left and start taking derivatives from the final token onwards. We will denote these derivatives for strings read right to left as $\partial^r_{w_i}$. Now we define the derivative of a language with respect to an alphabet element $w_i \in A$ as the set of all w_i prefixes; $\forall\, s \in \partial^r_{w_i} L, \text{concat}(s, w_i) \in L$. As in (4) the requirement for language membership can still be written as:

$$s = w_1 \cdots w_n \in L \text{ iff } \epsilon \in \partial^r_{w_1} \partial^r_{w_2} \cdots \partial^r_{w_n} L. \tag{5}$$

Now we will proceed to describing the connection between this parsing method and the training of autoregressive language models.

2.1.2 From membership proofs to the objective function

Autoregressive language modeling begins by constructing a map between the alphabet A and a vector space $\mathcal{E} = \mathbb{R}^{|A|}$ where $|A|$ is the cardinality of A. This mapping is such that every element of the alphabet (which in practice are sub-word tokens) is identified with a basis element in $\mathcal{E}$. Consequently, a string of length n is mapped to an element of $\mathcal{E}^n$, which is the n-fold cartesian product of $\mathcal{E}$. These mappings of tokens to basis vectors are referred to as 'one-hot' representations. Autoregressive language models are maps:

$$f_{\{\theta_I\}} : \mathcal{E}^n \to \mathcal{E}, \tag{6}$$

where $\{\theta_I\}$ are the neural network parameters where I is a multi-index running through indices of different components (layers) of the network as well as intra layer components of the network parameters. Training is the procedure by which $\{\theta_I\}$ are obtained via gradient descent on an objective function. The objective function is a map:

$$\mathcal{L}_{\{\theta_I\}} : \mathcal{E}^n \times \mathcal{E} \to \mathbb{R}, \tag{7}$$

i.e. a function that assigns a scalar to every input-target pair of the language model, again parameterized by $\{\theta_I\}$. The trained network parameters are obtained as:

$$\{\theta_I^*\} = \min_{\{\theta_I\}} \mathcal{L}_{\{\theta_I\}}(\vec{s}, \hat{t}) \tag{8}$$

where $\vec{s} \in \mathcal{E}^n$ and $\hat{t} \in \mathcal{E}$ are the one-hot representations of the input string and the target token. The specific form of this objective function used most commonly in language modeling is the so-called cross-entropy loss. Explicitly, this function reads:

$$\mathcal{L}_{\{\theta_I\}}(\vec{s}, \hat{t}) = -\sum_{a=1}^{|A|} \left(\log\left(\text{softmax}(f_{\{\theta_I\}}(\vec{s}))\right)^a \hat{t}^a\right). \tag{9}$$

Note that here we denote the loss for a particular input-output pair. The total loss is a sum of such terms for every input-target pair. Here softmax is a map from functions to normalized probabilities:

$$\text{softmax}(f_{\{\theta_I\}}(\vec{s}))^a = \frac{\exp(f^a_{\{\theta_I\}}(\vec{s}))}{\sum_{b=1}^{|A|} \exp(f^b_{\{\theta_I\}}(\vec{s}))}. \tag{10}$$

Notice that the expression for the cross-entropy is the negative log likelihood that the distribution given by the softmax of the network function equals the target distribution. The crucial fact to note about this function is that it is minimized when $f_{\{\theta_I\}}(\vec{s}) = \hat{t}$. In other words, minimizing this objective enforces the predictions of the model to match the targets.

The question then is how to choose the inputs and targets. Here we see the connection to parsing with derivatives. At each training step, we choose $(\vec{s}, \hat{t})$ s.t $s \in \partial_t^r L$ where L is the set of strings in the training corpus. In other

words, the inputs for the next token prediction are elements of the derivative of the language defined by the training corpus with respect to the targets, which are the desired tokens to be predicted! Furthermore, during the pre-training phase, for every string in the training corpus, we demonstrate the parsing by derivatives (from the right) to the language model. In other words, given a string $s = w_1 \cdots w_n$ we form from it the examples:

$$\left\{(s_1 \in \partial^r_{w_n} L, w_n), (s_2 \in \partial^r_{w_{n-1}} \partial^r_{w_n}) L, w_{n-1}, \cdots, (s_{n-1} \in \partial^r_{w_2} \partial^r_{w_{n-1}} \partial^r_{w_n}) L, w_2)\right\}. \tag{11}$$

Per string, the loss is given by:

$$\mathcal{L}_{\text{per string}} = \frac{1}{n}\left(\mathcal{L}_{\{\theta_I\}}(\hat{s}_1, \hat{t}_{w_n}) + \cdots + \mathcal{L}_{\{\theta_I\}}(\hat{s}_{n-1}, \hat{t}_{w_2})\right). \tag{12}$$

In each tuple above, the left-hand entry is the input and the right-hand entry is the target. Therefore, we train autoregressive language models by *enforcing* a parsing by derivatives from the right of the training corpus.

For a given string $s = w_1 \cdots w_n$, the expoential of the language modeling losses evaluated on all the next token examples the string yields is known as the perplexity:

$$p = 2^{\frac{1}{n}\left(\mathcal{L}_{\{\theta_I\}}(\vec{s}_1, \hat{t}_{w_n}) + \cdots + \mathcal{L}_{\{\theta_I\}}(\vec{s}_{n-1}, \hat{t}_{w_2})\right)}, \tag{13}$$

where $(\vec{s}_1, \hat{t}_{w_n}), ..., (\vec{s}_{n-1}, \hat{t}_{w_2})$ correspond to the vocabulary vector space representations of the examples formed from the string. When this measure is smaller than the exponential of the validation loss of the trained language model, then the language model "accepts" the string as a member of the natural language it was trained on. In this way, we can use trained language models to determine 'approximate' membership of strings to natural languages.

2.1.3 Neural scaling laws

The Neural scaling laws are an empirical relationship between the next-token prediction loss of transformer models on unseen data and 1) the number of parameters in the language model, 2) number of data points on which the model is trained and 3) the amount of computation measured in terms

of floating point operations applied to training the language model. The relationship between the loss and the number of parameters and datapoints takes the form:

$$\mathcal{L}(N,T) = \left(\left(\frac{N_c}{N} \right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D_c}{D} \right)^{\alpha_D}, \tag{14}$$

here N, D denote the number of (non embedding) parameters and the number of datapoints, and α_N, α_D, are constants that measure the effect of the gross number of parameters and datapoints, and N_c, D_c measure the critical thresholds around which either the model or dataset are too small for the loss to appreciably descrease (i.e. along a power law). The specific values of these constants depend on particulars of the model architecture, the dataset, and the training procedure.

For the purpose of our discussion, this phenomenological relation predicts that larger models trained on more data will get ever better at parsing natural language. We should note however that what these relations do not predict as well are the performance of the models on multi-choice benchmarks that test capabilities such as reasoning. This is to say that the downstream capabilities of language models are hard to infer directly from knowing how well they can furnish memberhip proofs of string in natural languages.

2.2 Parallels with Human Language Acquisition

Humans learn language by being exposed to linguistic input and rapidly developing an understanding of which words and sentences are acceptable within their language. Although the detail is very different than how LLMs learn language, both involve recognizing grammatical structures, syntax, and the meanings of words in various contexts. Although it may not be central to the learning process, humans intuitively predict how sentences should be completed, anticipate responses in conversations, and detect anomalies in language use. LLMs, through next-token prediction, perform analogous tasks by generating likely continuations of text based on learned patterns. Indeed, the psycholinguistics theory on prediction mechanism proposes a very similar mechanism of language acquisition in humans (cf. Ryskin and Nieuwland

(2023)). The idea here is that, during language processing, comprehenders predict upcoming linguistic input. These predictions draw on many sources of information including the preceding sentence context.

Chomskyan linguists claim children are born with an innate "Universal Grammar" to learn languages with minimal input (Chomsky, 1957). The apparent success of large language models (LLMs) in acquiring syntax without innate grammar challenges this view, as argued by Piantadosi (2023). LLMs show that statistical learners can induce syntactic rules, although they typically receive far more input than children and operate in distinct environments. Recent efforts, such as the BabyLM challenge, demonstrate that smaller models trained on child-directed data can efficiently learn grammar, suggesting that statistical models can learn grammar more effectively than previously thought. However, it remains an open question whether statistical learners without innate parsers can match the efficiency of children learning from limited input. Ongoing research aims to replicate children's learning environments more closely, using developmentally plausible spoken text or egocentric audiovisual data. If these models replicate the syntactic generalizations of children, it could further challenge the necessity of innate grammar for language acquisition. The question has clearly become one of efficiency, as opposed to one of principle.

One of the 'in principle' arguments of Chomsky (2023) is that LLMs could have learned 'impossible languages' and thus are bound to be semantic zombies. Again, recent experiments challenge this view, showing that compressibility of the input for LLM's is crucial for efficient learning, much like in humans. For instance, Tseng et al. (2024) investigates how large language models (LLMs) compress semantic pairs, finding that correct pairings result in better compression. Using semantic relations from English and Chinese Wordnet, the study shows that LLMs have an advantage in compressing texts with accurate semantic relations, measured by the compression advantages index. Larger models and those fine-tuned with structured knowledge (e.g., Chinese Wordnet) perform better, with fine-tuning greatly enhancing compression and semantic task performance. Notably, these improvements generalize to new tasks, suggesting the learning of abstract semantic con-

straints.

3 On Parrots, Octopuses, and Zombies

3.1 The Sensorimotor Grounding Argument

A common criticism of LLMs is that, lacking sensorimotor grounding, they cannot genuinely understand language or meaning. The idea is that because AIs do not interact with the world through senses and actions as humans do, they are merely manipulating symbols without grasping their semantic content; like a parrot who repeats the word 'sword' without it ever having seen or wielded one. Current AI's are, according to this criticism, "semantic zombies." This is one of the most common criticisms of LLM's, often assumed as almost a priori.

First, note that phenomenal (or qualia) zombies are logically possible only if one assumes that the qualia have no irreplaceable functional role, i.e., that they are functionally inert. Clearly one cannot argue for the existence of a semantic zombie in the same way without begging the question, since mental representations are typically postulated precisely for explaining behavior. So more needs to be said about what constitutes a semantic zombie.

Bender and Koller (2020) present a parable that is supposed to illustrate the kind of semantic vacuum that arises in the absence of direct acquaintance or sensorimotor grounding. The parable involves an eavesdropping octopus and is a close cousin of Searle (1980)'s Chinese room. It goes as follows. Two human speakers are stranded on different desert islands, but they can communicate with each other via an underwater cable. A hyper-intelligent octopus, with no knowledge of the surface world, wiretaps the cable.[1] By lis-

[1]The octopus is often used as a symbol for alien intelligence—intelligent but fundamentally different from human cognition. Octopuses have a distributed nervous system, a decentralized form of intelligence that challenges our anthropocentric view of mind and thought. In AI, the octopus can symbolize the potential for AI to evolve into a form of intelligence that is non-human and radically different in structure and operation. It suggests that AI may develop novel forms of reasoning or consciousness that don't mirror human thought processes, much like the octopus's mind is distinct from ours.

tening to the human conversation, the octopus learns to predict the speakers' responses with increasing accuracy. Bender and Koller claim that no matter how accurate the prediction, the octopus will never be able to grasp the meaning of words and sentences, since it has never been to the world in which they find their reference. In other words, the octopus has no true *knowledge* of the surface world. This thought-experiment is supposed to illustrate how, *pace* a successful symbol manipulation by the octopus, the absence of direct causal relations strips terms of reference from their intended meaning, and so makes *knowledge* about the world to which the words refer impossible.

3.2 Indirect grounding: where are we now?

This dichotomy between symbol manipulation and knowledge is reflected in much of the literature on AI today. The question is often put as follows: are LLMs better understood as tracking the referents of the words they use and (doing something isomorphic to) tracking and dealing with those referents, or are they better understood as tracking relationship between words? Which depending on how it is answered, leads to the idea that stochastic parroting can accurately describe the outputs, but it does not predict that LLMs have internal structure representing concepts that generalize, and do not expect that changes to one token affect different but conceptually related tokens.

Here we will reject this dichotomy. The texts on which LLM's have been trained are not random, but highly structured. More importantly, we, human beings, through centuries of writing, experimenting and researching, have encoded an immense amount of knowledge in these texts. The texts *are* tightly tethered to the world, and indeed, as argued in the previous Section, this is an important reason why LLM's are able to learn languages so effectively. And so meaning can arise with only indirect reference; indeed we often see this in abstract concepts and terms with no concrete referents.

Thus far, the argument is not novel. Lyre (2024); Piantadosi and Hill (2022) both argue that although LLMs do not have direct sensory or motor interactions with the world, the text data they are trained on is produced by humans who do. **?** distinguishes between inferential semantics (relation-

ships between expressions) and referential semantics (relationships between expressions and referents). He then argues that while Transformers excel in inferential semantics, they can also achieve referential semantics, precisely because they *are grounded* in representations of the physical, mental, and social world. He then evinces empirical evidence showing that language model vector spaces are near-isomorphic to brain imaging, perceptual, and physical spaces.

Indeed, recent evidence supports the idea that LLMs are developing world models that bear a notion of isomorphism to the trained data, and that they rely on the world model to generate sequences. Thus far, the evidence is for very simple concepts in an LLM's internal representations, such as color (Abdou et al., 2021), direction (Patel and Pavlick, 2022), etc. Nonetheless, they found that the representations for different classes of these concepts are easier to separate compared to those from randomly-initialized models. By comparing probe accuracies from trained language models with the probe accuracies from randomly-initialized baseline, they conclude that the language models are at least picking up something about these properties.

Similarly, in Li et al. (2022), this is explored for an LLM trained on the legal moves of the game Othello. Unlike reinforcement learning models like AlphaGo (Silver et al., 2016), which incorporate game rules and board structures to predict optimal moves, this model treats game sequences as generated text without explicit knowledge of board structure or rules. Othello-GPT learns only from lists of moves (e.g., E3, D3, C4) to predict the next move. Nonetheless, the trained model achieves a legal move accuracy of 99.99%, compared to 6.71% for an untrained version. They found not only that training induces an emergent spatial representation similar to the Othello board, but that it is possible to directly track interventions on the internal representation to the outputs of the model.

Here we want to propose a slightly different argument, which bypasses the notion of semantic grounding and understanding. In particular, it avoids the agent-centered character of so many of these concepts.[2] We now turn to

[2]For instance, in the context of scientific understanding, a recent review (Barman et al.,

this.

3.3 An argument based on knowledge and evidence

In Section 3.1 we equated the putative semantic hollowness of the octopus's propositions with a lack of knowledge, in particular about the terms used in those propositions, but more broadly about the surface world in which they originated.

Here, we define knowledge in the usual, minimalist, tripartite manner as 'justified true beliefs'. The definition can be understood within any theory of truth, and whether justification is seen as external (e.g. in terms of reliabilism or knowledge-first externalism) or internal (e.g. based on rational belief-formation rules). 'Beliefs' can also be understood in several ways. They can, for instance, be functionally characterised by consistency of responses to the same queries: if under repeated questioning, an LLM denies that Hilary Clinton won the 2016 election for president, we can say it believes the proposition that 'Hilary Clinton didn't win the 2016 US election for president'. Beliefs can also be seen externally, in physicalist terms, as a distribution of weights in a neural network, etc. Any such notion of 'belief' is sufficient for our purposes; we need not mention 'subjective feelings in our internal world'.

Indeed, none of what follows depends on which definitions of 'justification', 'true', and 'belief' we choose among those defended in the current literature, as long as we can agree that non-human cognizing agents are not logically barred from having knowledge. To the sceptic, functionalist definitions may seem more palatable, and, unlike in the case of phenomenal zombies, they do not beg the question against the semantic zombie (see

2024) proposes that:

> Understanding is an ability to "provide explanations within a theoretical framework that is intelligible to the agent, which involves the ability to derive qualitative results, answer questions, solve problems properly, and extend knowledge to other domains or levels of abstraction".

Lyre (2024) for a more detailed argument for functionalist responses to the semantic zombie).

In the context of AI, this shift from semantic grounding to knowledge is not uncommon (see e.g. Liu (2023) and references therein). For instance, in a paper about semantic grounding in AI, Biu (ibid) defines grounding as:

> the process of connecting abstract knowledge and natural language to the internal representations of our sensorimotor experiences in the real world and our subjective feelings in our internal world.

(We will deal with the role (or lack thereof) of subjective feelings shortly.) And the core of the claims of Section 3.1 about the primacy of sensorimotor experience for semantic grounding translates without loss to the context of knowledge and evidence. The idea is that only certain kinds of knowledge can serve as evidence. Evidence, in this critical view, must be directly grounded on experience, usually taken to be perceptual or sensorimotor. (Maher, 1996, p. 158) writes: "Even if a proposition is known to be true, if this knowledge [E] is not directly based on experience then E is not evidence and hence not evidence for anything."

But more recently, a less agentic notion of knowledge has been defended by Bird (2023). On Bird's view, there is no indispensable role for human sense-perception in the scientific process. Every role traditionally taken on by the human senses could be taken over by a reliable automated process without undermining the epistemic credentials of the output. 'It requires no great leap of imagination', says Bird, 'to see such robotic science becoming sufficiently reliable and routine that it is produced, published, and even consumed with minimal human intervention' (p. 93).

As argued convincingly by Bird, evidence is just knowledge that is used in successful inferences toward more knowledge. So the concept of evidence functionally characterises propositions by their role in inference. If a successful inference to further knowledge was based on some proposition E, then E constitutes knowledge, and if an inference relies on an unjustified premise, the conclusion will fail to count as knowledge. As illustrated by Bird (ibid,

p. 122):

> [...]intermediate propositions in the chain of inferences have the status of evidence propositions. The picture [...] is this. A scientist makes an inference that generates a conclusion. Then that scientist or another scientist uses that first conclusion to make another inference to some second, further conclusion. [...] if the inference to the second conclusion is knowledge-generating then the first conclusion is evidence. For example, Tycho Brahe made observations of the planets. From this evidence, Kepler inferred that planets travel in ellipses (and his other laws). Starting from Kepler's conclusions, Newton inferred a further conclusion, that the planets are subject to an inverse square central force law. [this argument] claims that Kepler's law that planets travel in ellipses is evidence—assuming that Newton came to know the force law as a result of his inference. The counterproposal [...] limits evidence propositions to non-inferential knowledge. [...] it denies that the inferred propositions, such as Kepler's laws, are evidence.

In order to be terminologically neutral, let us momentarily call whatever input AI's have provided humans as 'information'. And let us take as a given that the information that AI's have provided humans, whether in the context of chess or of protein-folding, has decidedly informed further development in these areas by humans. In formal jargon, that information has served as 'evidence' for further human knowledge formation.

It is at this point in time undeniable that current AI systems have produced 'knowledge' in the everyday sense of the word. Strategies for board games, such as GO or Chess have progressed since, and due to, the intervention of AI's. Or, in the real world, knowledge about protein-folding has undergone a similar advance. It is not unthinkable that an LLM trained on Othello (as described above) would provide something similar.

Thus, since (some of) the information provided by AIs can be characterised as a kind of belief that decidedly served as an inferential basis for

developing more knowledge, not only by us, but *by the LLMs*, it is itself knowledge.

The unique feature of this conception of knowledge is that it downplays the 'agentic' connotation of 'beliefs' in the tripartite (JTB) definition of knowledge. Moreover, as Bird explains, the very relational structure between a set of beliefs is what encodes knowledge (ibid, p. 129) :

> The counterfactual causal sensitivity of our beliefs to one another can be enough to ensure [that they reflect knowledge.] it is undoubtedly the case that it is typically the reliability (often causal) of the connection between the facts and a belief-like mental state that makes that mental state one of knowing.

Here, all we need to do to avail ourselves of this argument is to replace a human mental state, physically instantiated in a brain, by an LLM's state, physically instantiated by weights in a neural network. But of course, under externalism, there is no requirement to spell out what internal state justifies a belief. It is the external reliability of the belief-forming procedure that makes it knowledge. And there *is* a reliable causal connection between facts in the world and belief-like mental states of an LLM: we have supplied this connection ourselves in training and post-training. The counterfactual sensitivities of an LLM's beliefs to one another will reflect knowledge just as ours do.

The counterproposal is that only non-inferential knowledge can serve as evidence. Again, Bird illustrates (ibid p.123)

> This counterproposal still gives evidence a key epistemological role. This role is foundational. In the chain of inferences just considered, it is only the initial propositions that are the evidence propositions. The intermediate propositions do not count as evidence, according to the counterproposal. Evidence propositions are those propositions that form a basis of our inferences and are not themselves inferred from anything else.

The appeal of non-inferential knowledge, usually assumed to be a type of 'raw' perceptual, or sensorimotor information which AI's and LLM certainly

lack, is that it is often construed as certain, or nearly certain. However, there is no reason to think that non-inferential knowledge, or any knowledge, has this kind of certainty. Indeed, optical and sensorimotor illusions abound, and evidence from our senses often needs to be corrected by more precise scientific instruments. Although many lines of scientific instruments begin by aiding human perception, they quickly outgrow that function and end up replacing human perception altogether.

As Bird (2023) successfully argues, there is no reason for the concept of evidence to be constrained to be non-inferential. First off, we generally do not remember the non-inferential propositions that form the basis of our inferential knowledge. I can lose my sensorimotor or perceptual evidence for some proposition without also losing knowledge of what was successfuly inferred from it. Indeed, what is retained is usually the information extracted from the original experience, not the content of the experience itself. As Bird writes, (ibid, p 127) "I can remember the melting point of lead ($327.5 \pm C$), but not the visual or auditory experience by which I learned this fact." We are constantly acquiring new evidence for many of our existing beliefs. And we don't need to repeatedly re-confirm those beliefs through fresh inferences.

Summing up our response to Bender and Koller (2020)'s octopus' thought-experiment, we should ask: how is the octopus's access to the surface world qualitatively different from human access to the world of viruses and bacteria, not to mention quarks and gluons? We certainly do not directly perceive these with our unaided senses. More and more, what we directly perceive are only numbers and graphs on a computer screen. Nonetheless, we claim to have knowledge of these theoretical items. We rely on heavily theory-laden, iterative processes to lead us to justified true beliefs. Perhaps with enough time and assuming the stranded humans were reporting on all kinds of occurrences and structures of their surface-world—were reporting, that is, knowledge—the octopus just might know what is up.

3.4 Summing up

First, it is important to issue a caveat: this argument primarily applies to propositional, or conceptual knowledge, that can be expressed in declarative sentences. There may be forms of non-propositional or tacit knowledge, such as experiential or embodied understanding, that AIs do not possess. An LLM doesn't have the know-how to ride a bicycle, for example. While the scope of this kind of knowledge is a hotly debated topic (cf. Stanley (2011)), we concede that such forms of knowledge are not yet replicated in current LLMs, but we won't speculate on whether there are future avenues for doing so.

So, to sum up: philosophical arguments suggest that direct sensorimotor experience is not the sole basis for knowledge. In this viewpoint, knowledge consists of justified true beliefs, and evidence is just knowledge used in successful inferences to acquire further knowledge. This viewpoint implies that an entity can possess knowledge if it can process and infer information in ways that lead to true beliefs, regardless of the origin of its initial knowledge base.

In this section, we have challenged the notion that AIs are semantic zombies by arguing that sensorimotor grounding is not a prerequisite for possessing knowledge or engaging with meaning. By reframing knowledge in terms of justified true beliefs and functional inference, we open the possibility that AIs have cognitive capacities more akin to human understanding than critics suggest. Thus we proposed that:

1. **Inferring Justified True Beliefs:** AIs are capable of processing inputs and, through learned patterns, producing outputs that correspond to justified true beliefs. Their inferences, grounded in extensive data processed during training, allow them to generate knowledge that is coherent and applicable.

2. **Linking Knowledge to Meaning:** Meaning can be construed in terms of the ability to use information to make accurate inferences and predictions. If an AI can use linguistic inputs to generate valid conclusions, it engages with knowledge in a functionally similar manner

to humans.

3. **Defining Beliefs in AIs:** Beliefs can be considered as stored representations or states that guide behavior and inference. In AIs, the learned parameters and internal states serve this purpose, guiding the generation of outputs based on inputs.

By these points, we argued that AIs possess a form of knowledge grounded in human-provided data.

References

M. Abdou, A. Kulmizev, D. Hershcovich, S. Frank, E. Pavlick, and A. Søgaard. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online, 2021. Association for Computational Linguistics.

Kristian Gonzalez Barman, Sascha Caron, Tom Claassen, and Henk de Regt. Towards a Benchmark for Scientific Understanding in Humans and Machines. *Minds and Machines*, 34(1), April 2024. ISSN 1572-8641. 10.1007/s11023-024-09657-1.

E. M. Bender and A. Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Online, 2020. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM, March 2021. 10.1145/3442188.3445922.

Alexander Bird. *Knowing Science*. Oxford University Press, Incorporated, Oxford, 2023. ISBN 9780192606822. Description based on publisher supplied metadata and other sources.

David J. Chalmers. Does Thought Require Sensory Grounding? From Pure Thinkers to Large Language Models. *Proceedings and Addresses of the American Philosophical Association*, 97:22–45, 2023.

Noam Chomsky. *Syntactic Structures*. Mouton de Gruyter, s.l., 1. aufl. edition, 1957. ISBN 3110218321.

Noam Chomsky. The False Promise of Chatgpt. *The New York Times, March 8, 2023*, 2023.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*, 2022.

Bing Liu. Grounding for Artificial Intelligence. *arxiv:2312.09532v1*, 2023.

Holger Lyre. "understanding ai": Semantic grounding in large language models, 2024. URL `https://arxiv.org/abs/2402.10992`.

Patrick Maher. Subjective and objective confirmation. *Philosophy of Science*, 63(2):149–174, 1996.

Melanie Mitchell. Artificial Intelligence Hits the Barrier of Meaning. *Information*, 10(2):51, February 2019. ISSN 2078-2489. 10.3390/info10020051.

Melanie Mitchell. On Crashing the Barrier of Meaning in Artificial Intelligence. *AI Magazine*, 41(2):86–92, June 2020. ISSN 2371-9621. 10.1609/aimag.v41i2.5259.

Dimitri Coelho Mollo and Raphaël Millière. The vector grounding problem, 2023. URL `https://arxiv.org/abs/2304.01481`.

Roma Patel and Ellie Pavlick. Mapping Language Models to Grounded Conceptual Spaces. In *ICLR*. OpenReview.net, 2022. URL `http://dblp.uni-trier.de/db/conf/iclr/iclr2022.html#PatelP22`.

Ellie Pavlick. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251), June 2023. ISSN 1471-2962. 10.1098/rsta.2022.0041.

Steven T. Piantadosi. Modern language models refute Chomsky's approach to language. 2023.

Steven T. Piantadosi and Felix Hill. Meaning without reference in large language models. *arxiv: 2208.02957*, 2022. URL `https://arxiv.org/abs/2208.02957`.

Rachel Ryskin and Mante S. Nieuwland. Prediction during language comprehension: what is next? *Trends in Cognitive Sciences*, 27(11):1032–1052, November 2023. ISSN 1364-6613. 10.1016/j.tics.2023.08.003.

John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424, September 1980. ISSN 1469-1825. 10.1017/s0140525x00005756.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016. ISSN 1476-4687. 10.1038/nature16961.

Jason Stanley. *Know How*. Oxford University Press, August 2011. ISBN 9780199695362. 10.1093/acprof:oso/9780199695362.001.0001.

Anders Søgaard. Understanding models understanding language. *Synthese*, 200(6), October 2022. ISSN 1573-0964. 10.1007/s11229-022-03931-4.

Anders Søgaard. Grounding the Vector Space of an Octopus: Word Meaning from Raw Text. *Minds and Machines*, 33(1):33–54, January 2023. ISSN 1572-8641. 10.1007/s11023-023-09622-4.

Yu-Hsiang Tseng, Pin-Er Chen, Da-Chen Lian, and Shu-Kai Hsieh. The Semantic Relations in LLMs: An Information-theoretic Compression Approach. In *NEUSYMBRIDGE*, 2024. URL `https://api.semanticscholar.org/CorpusID:269950935`.