Opinion

From task structures to world models: what do LLMs know?

Ilker Yildirim1,3,4,5,* and L.A. Paul2,4,6,*

In what sense does a large language model (LLM) have knowledge? We answer by granting LLMs ‘instrumental knowledge’: knowledge gained by using next-word generation as an instrument. We then ask how instrumental knowledge is related to the ordinary, ‘worldly knowledge’ exhibited by humans, and explore this question in terms of the degree to which instrumental knowledge can be said to incorporate the structured world models of cognitive science. We discuss ways LLMs could recover degrees of worldly knowledge and suggest that such recovery will be governed by an implicit, resource-rational tradeoff between world models and tasks. Our answer to this question extends beyond the capabilities of a particular AI system and challenges assumptions about the nature of knowledge and intelligence.

Kinds of knowledge

OpenAI’s GPT-4 [1] and other LLMs, such as Meta’s LLaMA [2], show impressive conversational capabilities. These systems can generate coherent, novel, and often surprisingly sophisticated responses to questions or prompts posed directly in natural language. This has led artificial intelligence (AI) researchers who develop these systems, along with cognitive scientists and authors of articles in news media, to ask whether LLMs have knowledge.

Here, we explore this possibility. We suggest we are in a ‘Kuhnian moment’: a conceptual revolution in what we take knowledge to involve, with implications for how we think intelligence could arise. Accordingly, we ask: in what sense can GPT-4 (and similar models) be said to have knowledge? The answer to this question extends far beyond the capabilities of a particular AI chatbot, with implications for cognitive science, neuroscience, philosophy, and AI.

We ground our answer using a core concept from cognitive science, world models (see Glossary). World models are structure-preserving, behaviorally efficacious representations of the entities and processes in the real world [3], including objects with 3D shapes and physical properties [4], scenes with topological relations and navigable surfaces [5], and agents with beliefs and desires [6]. Human thought often relies on these types of world models [7,8] to perceive [4,9], effectively reason [10,11], plan [12], and talk about the world [13].

In particular, for a wide range of ordinary contexts, a knowledgeable human agent draws on their world model, exploiting a structural match between their mental representations and the state of the world and, using language, reliably generating contentful answers to prompts that are approximately truth preserving and relevant. We describe such world-model-based knowledge as worldly knowledge and the content of the matched representations that support it as worldly content. When a subject has worldly knowledge of a proposition, this is, at least in part, in virtue of the subject using their world model to grasp the worldly content of the proposition. For example, when the subject knows that balancing a ball on a box is easier than balancing a box on a ball, they use their world

Highlights

OpenAI’s GPT-4 and similar large language models (LLMs) show impressive conversational capabilities. This Opinion asks: in what sense does a LLM have knowledge? The answer to this question extends beyond the capabilities of a particular AI chatbot and challenges our assumptions about the nature of knowledge and intelligence.

We answer by granting LLMs ‘instrumental knowledge’: knowledge defined by a certain set of abilities.

How is such knowledge related to the more ordinary, ‘worldly’ knowledge exhibited by humans? To address this, we turn to a core concept in cognitive science, world models, and explore the degree to which instrumental knowledge might incorporate such structured representations.

We discuss how LLMs could recover degrees of worldly knowledge and suggest that such recovery will be governed by an implicit, resource-rational tradeoff between world models and task demands.
model of how objects move and react to external forces to grasp the content that balancing a ball on a box is easier than balancing a box on a ball. We take such knowledge to be the target of philosophical analyses of how an individual knows the content of a proposition \( p \), where to know that \( p \), ‘an agent must not only have the mental state of believing that \( p \), but various further independent conditions must also be met; \( p \) must be true, the agent’s belief in \( p \) must be justified or well-founded, and so forth’ ([14], p. 281). We also take such knowledge to be well studied in scientific contexts [15–20] and to include much of our ordinary factual and relational knowledge about the world.

Recent AI developments in LLMs seem to involve a different kind of knowledge. Such systems are based on deep neural networks pretrained on Internet-scale data to autocomplete the next word (or token, more accurately) given preceding context, which are then further fine-tuned with reinforcement and supervised learning techniques for human-aligned and humanlike responses (e.g., [21]). When using these tools, LLMs can generate sufficiently successful responses to an appropriately wide range of prompts, giving answers that are often approximately truth preserving and relevant. We describe such answers as demonstrating instrumental knowledge.

Instrumental knowledge is knowledge acquired through the successful use of instruments that perform certain tasks. A person who can successfully use an instrument like a television remote gains a small amount of instrumental knowledge, as evidenced by the limited range of tasks the remote can be used to perform. (The person gains this instrumental knowledge through task performance, not through knowing in any deeper sense how the television remote works, making their knowledge merely ‘instrumental’.) Here, we treat the process of next-word generation as an instrument (cf. [22] for a treatment of language as an instrument). We take an LLM’s wide-ranging and generative ability to successfully respond to prompts, using next-word generation, as demonstrating instrumental knowledge. An LLM that can successfully use the instrument of next-word generation gains a significant amount of instrumental knowledge meeting a high standard of reliability, as evidenced by the range of tasks it can use this language tool to successfully perform. Note that the LLM gains its instrumental knowledge through task performance, not through knowing, in any deeper sense, how its responses are about the world.

So, we can grant that LLMs exhibit knowledge, as, like people, they can be said to have instrumental knowledge. However, how is such instrumental knowledge related to the sort of human knowledge that is based on world models and to what degree, if any, might an LLM’s knowledge incorporate worldly knowledge?

World models in cognition

We start with an exposition of world models as realizers of worldly knowledge. Our definition of world models is inspired by Gallistel and King’s discussion of mental representations [3] and is partially grounded in mathematical representation theory. A world model is a way of representing entities in the real world and their relations with two critical requirements.

First, a world model must be structure preserving, such that changes in the real-world entities and their relations should map onto similar sorts of changes in their counterpart representations in the world model. Gallistel and King use the simple example of measuring a child’s height by marking it on a wall with a pencil, where the markings and the process of making these markings constitute a structure-preserving representation of the child’s height. Notice that a structure-preserving representation is, by definition, content preserving.

Second, a world model must be behaviorally efficacious, meaning that it should enable accurate planning and high-reward actions back in the real world. In the example above, the markings are
behaviorally efficacious as they can guide decisions about clothing size or comparisons of the heights of a child across time. Notice that behavioral efficacy suggests that world models need not be veridical replicas of what is out there; abstractions that can be implemented in algorithmically efficient approximations of their counterpart real world processes will do. Examples of such structure-preserving, behaviorally efficacious representations are provided in Figure 1 in Box 1.

Multiple lines of research provide support for world models as a basis of cognition. Probabilistic models in cognitive science suggest concrete candidates for how these structure-preserving, behaviorally efficacious representations may be implemented in the mind (see Figure 1 in Box 1). Embedding a candidate representation within probabilistic models, often referred to as causal generative models, allows researchers, in a given domain to formally and rigorously specify queries of inference, learning, prediction, and planning and to solve them using Bayesian inference. These models are typically evaluated against alternative models in how consistent they are with human performance, with respect to average accuracy, stimulus-driven variability in accuracy, response times, similarity judgments, and other sorts of behavioral ratings.

Beyond evaluating these models in comparison with behavioral performance, recent neuroscientific studies explore the realism of the hypothesized world models in brain activity. A formal review of this literature is beyond the scope of this Opinion, but we provide brief pointers in Box 1. We now turn to the specification of instrumental knowledge in LLMs and situate this knowledge relative to worldly knowledge as characterized using world models.

**Instrumental knowledge**

How can we characterize the instrumental knowledge of LLMs? We can understand the instrumental knowledge of an entity in terms of its ability to use an instrument to perform tasks posed for it across relevant domains. Indeed, a motivating perspective in LLMs is the idea of unsupervised multitask learning [23]. Internet-scale natural language data can be seen as a large dataset of a multitude of tasks posed in varying ways and forms, consistent with the messiness of how language is used naturally. For instance, the abbreviation ‘TL;DR’ or a paragraph that starts with the phrase ‘In summary,…’ might signal a summarization task; nearby or paired sentences or phrases spanning multiple languages might suggest the task of translation between those languages. For a model to accurately predict the next word in a sequence, researchers have speculated that it may be critical for the model to spontaneously infer the task structure from the preceding context and condition the next-word predictions on that task structure [23]. After the training of an LLM is over, using next-word generation to infer such task structure from natural language, and conditioning the activations within the model according to this structure, is a possible source of instrumental knowledge (Figure 1A,B).

Could inference and use of task structure occur without (or with very little) worldly knowledge? A setting where this could occur is machine language translation. Instead of focusing on building systems that translate through semantic analyzers or any other formal notion of meaning, most progress in machine translation relies on increasingly sophisticated statistical approaches [24,25]. It is plausible that LLMs represent a new frontier in this progression of models, one in which the models infer the task structure of language translation in terms of how words, phrases, and even paragraphs are emitted within and across pairs of languages, and use this structure to translate – without necessarily representing worldly knowledge or rendering it in different languages. Such a possibility is further suggested by ‘relational’ theories of word meaning (e.g., a conceptual role semantics where the meaning of a word or phrase is defined directly by its relation to other words or concepts and only indirectly through reference and causal connections to the nonlinguistic world, with limited transmission of worldly content) [26].
Could inferring and using task structure be reduced to having a grasp of the rules and patterns of a language? Some have asked whether LLM ‘knowledge’ is merely an ability to follow linguistic rules and language patterns. In addressing this question, an important distinction concerning the performance of LLMs is between ‘form’ versus ‘meaning’. As others have rightfully cautioned [27], an LLM’s ability to generate coherent language should not be taken as evidence of
understanding natural language. (In this Opinion, we relate such understanding to having worldly knowledge.) Similarly, based on the separation observed in the human brain between language and non-language regions [28,29], researchers have argued [30] that LLMs acquire formal linguistic competence, or knowledge of the rules [30] and statistical regularities [31] of a language, but not ‘functional linguistic competence’, which includes knowledge of and reference to things and processes in the social and physical worlds [32]. (In this Opinion, we see functional linguistic competence as related to worldly knowledge.)

When making this argument, researchers have sometimes drawn a distinction between pretrained LLMs (on next-word prediction) and LLMs after fine-tuning with supervised or reinforcement learning objectives on human dialog data (e.g., [21]). When prompted with examples that seem to require worldly knowledge, some fine-tuned LLMs, including GPT-4, generate compelling answers [21]. It is of interest to understand how such fine-tuning impacts knowledge; however, such procedures typically adapt only certain output stages of the pretrained models, otherwise keeping much of the pretrained weights frozen [33,34]. Moreover, in certain cases, performance similar to that of a fine-tuned model can be obtained with pretrained LLMs via so-called ‘in-context learning’; in the absence of any parameter updates in the underlying model, providing a few example input–output pairs in the prompt can lead
LLMs to learn new tasks [35]. We see in-context learning as relying on instrumental knowledge that exceeds ‘formal linguistic competence’ and includes inference of a task structure and conditioning of next-word prediction on that structure (Figure 1A,B).

**The leap: from next-word prediction to world models**

Beyond inferring task structure, can the instrument of next-word generation also lead to inference of world models? In other words, could this purely text-related instrument lead to any degree of worldly knowledge (Figure 1C)?

The ordinary sense of ‘knowing a proposition $p$’, where $p$ is about the world, includes some degree of worldly content. One interesting way this could occur in an LLM draws on the ability to recover worldly content using **compression**. In many ways, next-word prediction in LLMs reflects compression of the vast amounts of text data crawled on the Internet into the many billions of weights of a deep neural network, which in proportion remains too small to memorize the training data. Indeed, compression and prediction are closely related objectives [36,37].

A lower-dimensional state space factorizing the relevant dimensions of variation of a given domain and the dependence of these dimensions on each other can be simultaneously used to compress and predict. World models are examples of such state spaces, where a small set of variables captures causal abstractions of the structure of their counterpart physical processes in the real world (see Figure 1 in Box 1). A caveat, however, is that compression does not always preserve worldly content, as when various possible approximations of the statistical regularities in the training text are not structure-preserving representations.

However, we speculate that it is possible for compression to recover a **structure-preserving mapping** of the data-generating process underlying the training data. There are multiple data-generating processes underlying natural language data that could be the source for this: the rules and statistical regularities in a language, the tasks that are posed and addressed in natural language, and, especially relevant to the present context, worldly content involving entities, physical processes, and situations projected into text by humans perceiving, talking about, and participating in these situations.

We acknowledge that, at present, to hold that LLMs do or even could recover structure-preserving abstractions of the world involves a leap of faith. However, as it is often this type of thesis that motivates attributions of general intelligence to LLMs (e.g., [38]), we turn to available literature exploring this possibility.

**Measured recovery of world models in LLMs under domain-specific settings**

World models (i.e., causal abstractions of a represented system with a structure-preserving representing system [3]) provide a concrete framework to reason about whether and how LLMs can recover degrees of worldly content that could lead to worldly knowledge. A small number of recent studies have explored whether a language model trained to predict next-token sequence spontaneously approximates the underlying data generating process (i.e., the world model [39,40]). We consider two lines of work: models trained on specialized, non-language domains and models trained on Internet-scale natural language data.

**Models trained on specialized non-language domains**

A set of recent studies analyzed models after training them on sequences of word-like tokens but in specific non-language domains where the unique tokens, as well as how they combine into sequences, are constrained by the underlying world model.
One example of the recovery of worldly knowledge is presented by Li et al. [40] (Figure 2B). This study trained a GPT-4-like (but on a much smaller scale) language model to predict legal moves in the game of Othello—a two-player board game in which players gain more discs by outflanking the discs of the other player. The world model in this game comprises the state of the board (for each cell in an $8 \times 8$ grid, whether it is black, white, or empty) and the set of rules by which each player changes this state. They randomly simulated this world model creating a dataset of board-state traces via stochastic, non-strategic decisions for each player. When they trained a language model, called Othello-GPT, on this dataset, they found that not only could this model reliably generate legal moves given the prior set of moves, but also the entire board state could be accurately decoded from the intermediate-layer activations in the model, with a linear decoder (as established by a follow-up study [41]). Crucially, the authors also showed that intervening on the board state via these decoders causally and appropriately impacted the model’s legal move predictions.

A recent study [39], using a toy domain, took a similar approach to suggest that a language model trained for next-token prediction for program synthesis can recover something about the deeper semantics of this domain-specific programming language. Finally, a similar conclusion comes from the domain of computational biology; a recent LLM trained by Meta researchers [42] to predict masked entries in protein amino acid sequences rendered the coarse 3D structure (i.e., contact relations between amino acids) of the actual folded protein linearly decodable.

The work by Li et al. [40] engages with the possibility mentioned earlier: that next-token prediction with a language model can recover structure through the underlying data-generating process. It suggests that the ‘leap’ mentioned in the previous section is a realistic outcome. It is exciting and of pressing importance for future research to systematically explore this possibility across the dimensions of training objectives (e.g., next-token prediction, masked token prediction), network architectures (e.g., transformer-based language models [43] and RNN-based sequence models [44,45], as these are the kinds of neural network architectures that underlie the modern LLMs and traditional neural language models, respectively), and the complexity of world models.

**Figure 2.** Using world models to explore the extent of worldly knowledge in large language models (LLMs). (B) Compression and prediction are like the two sides of a coin, and it is possible that compression can recover the data-generating process. In a domain-specific setting (the board game Othello) with tokenized traces of randomly generated game states, the work in [40] trained an LLM on next-token prediction. Surprisingly, the intermediate layers of this LLM yielded a linearly decodable full board state. (C) As we discuss in the text, generalization of this result to actual natural-language-trained LLMs is so far limited.
That said, we raise three caveats. First, these studies depend on dense sampling of relatively small domains (more than 1 million training examples for Othello); dataset requirements can quickly grow for non-toy domains, and whether even the largest Internet-scale language datasets can satisfy these conditions is unknown. Second, these studies consider settings where the basic building blocks of the world models are enumerated and assigned unique tokens (e.g., the locations in the 8 x 8 grid and the possible states for each cell in the Othello environment). We suspect this will not be applicable to many relevant world models and to how they are projected into text by human language users. Finally, it is possible that the recovered world model in a language model, based on the objective of next-token prediction, is computationally not the most efficient one.

Models trained on Internet-scale natural language data
This motivates the exploration of the recovery of domain-specific world models in LLMs trained with natural language data (Figure 2C) [46–48], drawing on recent work in the context of perceptual color space [46,47].

The work by Abdou et al. [47] shows that pretrained language models can recover aspects of the relational structure of the perceptual color space. Using representational similarity analysis, the researchers report statistically significant correlations between the similarity structure of color pairs with respect to language model embeddings versus Euclidean distances between the same pairs of colors under a well-established perceptual color space. Another study [46] provides further evidence but with a different approach. The researchers report that LLMs provided with a small number of in-context examples from a single hue (e.g., red) generalize, more accurately than chance, to the rest of the color space, indicating that some aspects of the relational structure of the perceptual color space are readily available in these LLMs.

In a way, these results suggest the possibility of recovering worldly knowledge in LLMs despite their purely text-based training. However, this needs to be qualified, as the quantitative nature of the correspondence between the physical spaces and LLM internals is often underwhelming in these studies (e.g., a correlation value of roughly 0.2 in [47]). We anticipate that this correspondence will increase under better trained, larger models; nevertheless, the fact that this relationship is weak in a domain like colors is telling; structure-wise, color space is a simple topology (distances in 3D space) and presumably there is much text in the training corpora that talks about color. The complexity of typical world models projected to text by human language users (e.g., spatial structures, intuitive physics) is often far more complex, suggesting that there is a significant amount of worldly content that still needs to be captured.

Across the settings we reviewed, from Othello-GPT to the case of perceptual color space, LLMs’ knowledge can incorporate varying degrees of worldly knowledge, from linearly mappable to rather limited correspondence to the underlying data-generating processes. Next, we speculate on an account of this observed variability, in terms of the complexity of the set of world models underlying a dataset and the tasks a system needs to perform.

A resource-rational approach
What might determine the degree of worldly knowledge recovered due to next-word generation in an LLM? We suggest that recent research in cognitive science offers two key dimensions to consider: the distribution of tasks an agent encounters [49] and the granularity or complexity of representations needed to accomplish these tasks [50–52] (Figure 3A). The combination of task distribution and the coarseness (or the overall complexity) of the structure-preserving representations that this task distribution requires might render the convergence of compression, under next-word generation, to a structure-preserving representation more or less likely (Figure 3B).
Worldly knowledge can be computationally costly; some of the mappings illustrated in Box 1 (e.g., game-engine-style physics simulation, embodied planning), although they are domain specific, may rival LLMs in the complexity of the computations they involve, as they map to complex causal processes in the physical world. For next-word generation to rely on world models (or any other stimulus-computable system), there is the additional challenge of inferring specific configurations of world models from specific input stimulus: going from a sensory measurement to a specific world model configuration, or from a proposition such as balancing a box on a ball is hard to a model in which this proposition is reflected, is an underdetermined problem of an intractable nature.

Accordingly, any system with bounded resources (in computation, time, memory), whether it is an LLM or a human, should take advantage of the efficiencies afforded by the distribution of tasks in their environments [53]. To see an example where tasks can modulate whether coarser worldly knowledge would be sufficient, consider the domain of intuitive physics (see Figure 1, top in Box 1): A coarse-grained, qualitative simulation might suffice to predict whether a liquid will flow toward right or left, while a finer-grained simulation may be necessary to determine details of its trajectory [54,55]. Figure 3A illustrates another scenario in the context of a navigation-related task. It is possible that the distribution of tasks an agent encounters may require only coarse-grained representations; alternatively, it is also possible that these tasks require only a specific small subset of the world configurations to be inferred, rendering more idiosyncratic knowledge sufficient.

Instrumental knowledge, inference of task structure, and conditioning of next-word generation on that, without much worldly knowledge, will likely be a sufficient ‘shortcut’ under a wide range of task distributions. For instance, even in the Othello-GPT study reviewed earlier, when the model is trained on a subspace of tasks based on a distribution of strategic gameplay, as

Figure 3. Resource-rational view of world models and instrumental knowledge. (A) A goal-conditioned mapping would invest fewer resources (e.g., larger ‘error bars’) on the aspects of worldly knowledge not benefiting the task of the agent. Such a mapping can remain behaviorally efficacious and structure preserving. (B) The extent to which next-word generation may recover worldly knowledge might follow from the ‘coverage’ of the task distribution relative to the underlying data-generating process and the granularity or overall complexity of this data-generating process. We expect recovery of worldly knowledge to be more likely with coarser representations and broader task distributions. The density of the black grids indicates the granularity of the representation; pink regions indicate the subspace of the representations implicated by the task distribution.
Box 2. World models as programmable, mid- or high-level interfaces for safety and alignment

Beyond their role in intelligence, why else should we embrace the use of world models in an AI system? That is, if a system has instrumental knowledge allowing it to perform a diverse range of tasks accurately, what else should we want?

We want such systems to be safe. That is, we want these AI systems to be deployed in a way that is both truthful and aligned with ‘human values’ [76]. Existing deep neural networks, including LLMs with their transformer-based neural network architectures, are black-box systems without an explicit, high-level interface to direct program their behavior. The lack of transparency in how LLMs work, combined with their tendency to produce so-called ‘hallucinations’ – the frequently observed fabrication of situations, events, and persons by these systems in response to reasonable prompts – has raised concerns in the industry and in larger societal contexts [77,78]. The possibility of unexpected value change in these systems (e.g., transformative higher-order value change as detailed in [79]) raises questions about how sure we can be that evolving AI values will remain aligned with human values [78].

World models may help us find a path to safer, truthful, and better aligned AI systems. Why? Because world models, and more generally domain-specific high-level programming languages, formalize worldly knowledge in structure-preserving, interpretable representations, and they can readily enable truthfulness by supporting an engineer or a user who wishes to impart their ‘values’ and safety measures as explicit constraints over the system. These features can be exploited via hybrid pipelines of LLMs and world models [30,81], by neuro-symbolic architectures [83], and by the creation of natively programmable neural network architectures [83,84].

opposed to a dataset of randomly unfolding legal game states, the decodability of the game state from the neural network activations (even with nonlinear decoders) is dramatically reduced. Similarly, for the natural language settings discussed (e.g., color space), typical next-word generation queries constrain what knowledge an LLM arrives at.

It is only recently that the field of cognitive science has started exploring computational theories that examine structured world models with task demands (Figure 3A). For example, recent work provides computational-level explanations of simplified representations relative to navigation-related objectives, using the domain of 2D maze navigation [52] and 3D scene perception [56]. This new landscape of resource-rational world models could in turn help to refine what we take knowledge in LLMs and humans to be.

Concluding remarks

The impressive performance of LLMs on a surprisingly broad range of tasks challenges how we think about the acquisition of knowledge, the relationship between instrumental knowledge and worldly knowledge, and, by extension, how intelligence could arise in artificial or machine systems. To frame this challenge, we asked: in what sense can LLMs, trained purely on text, primarily to predict the next word, be said to have knowledge? We answered the challenge by granting ‘instrumental knowledge’ to LLMs: knowledge defined by a set of (sufficiently sophisticated) abilities to use next-word generation as an instrument, including spontaneous inference and use of task structure. We then asked how such knowledge is related to the more ordinary, ‘worldly’ knowledge that people exhibit, and explored the degree to which LLM knowledge could incorporate world models. We also suggested two resource-rational frameworks from cognitive science – inference networks and goal-conditioned world models – as promising formalisms for how instrumental and worldly knowledge can overlap or interface. We close by noting that, beyond their implication for intelligence, world models, incorporated more explicitly in AI systems, will facilitate a more direct path to safe and aligned deployment by exposing an interpretable mid- or high-level interface for control and intervention (Box 2; see Outstanding questions).

Acknowledgments

We thank Yutaro Yamada, Mario Belledonne, Jonathan Schaffer, the Editor, and two anonymous referees for helpful discussions. We thank members of the Cognitive and Neural Computation Laboratory at Yale for their feedback on earlier versions of the manuscript. We thank Aalap Shah for his help making display items. I.Y. was supported by AFOSR grant # FA9550-22-1-0041.

Outstanding questions

How can we formalize instrumental knowledge? One possibility is task-conditioned world models, which needs further development. Future work should also explore the relationship of instrumental knowledge to amortized inference or data-driven proposals in Bayesian inference and resource-rational solutions to intractable problems arising from work with expressive world models.

Under what conditions – with respect to the underlying data-generating process, actual training data, and model architecture – does next-word prediction lead to approximate recovery of world models?

What is the impact of fine-tuning on an LLM’s instrumental knowledge and recovery of worldly knowledge?

How can we create diagnostic, domain-specific benchmarks to assess the extent and nature of worldly knowledge in LLMs? These benchmarks should explore dimensions such as world model granularity and the training data distribution.

To what extent can the distinction of instrumental knowledge versus worldly knowledge help toward the building and exploration of new foundation models in language and beyond, such as computer vision and reinforcement learning? (LLMs are often referred to as ‘foundation models’, in the sense of their adaptability to new tasks with little additional data.)

What is the format of the formal linguistic competence internalized by LLMs, specifically in relation to linguistic theories?

How does our discussion relate to existing treatments of LLMs in related contexts, including conceptual role semantics and embodiment-based arguments?

How can we incorporate structured world models in AI, such as LLMs and beyond, for safer and better aligned systems?
Declaration of interests
No interests are declared.

Resources
www.youtube.com/watch?v=SjhIlw3Iffs
https://writings.stephenwolfram.com/2023/03/chatgpt-gets-its-wolfram-superpowers

References