Meaning without reference in large language models

Steven T. Piantadosi
Department of Psychology
Helen Wills Neuroscience Institute
University of California, Berkeley
stp@berkeley.edu

Felix Hill
DeepMind
felixhill@deepmind.com

Abstract

The widespread success of large language models (LLMs) has been met with skepticism that they possess anything like human concepts or meanings. Contrary to claims that LLMs possess no meaning whatsoever, we argue that they likely capture important aspects of meaning, and moreover work in a way that approximates a compelling account of human cognition in which meaning arises from *conceptual role*. Because conceptual role is defined by the relationships between internal representational states, meaning cannot be determined from a model's architecture, training data, or objective function, but only by examination of how its internal states relate to each other. This approach may clarify why and how LLMs are so successful and suggest how they can be made more human-like.

LARGE language models (LLMs) have begun to display an array of competencies that were long thought to be out of the reach of neural networks. At the same time, critics have been vocal that existing methods, model structures, and training paradigms will be insufficient for anything like human language use. In particular, it is argued that LLMs will never achieve "meaning" or "understanding" due to either the objective function they optimize, the format of their internal representations, or the type of training data that they receive. This short comment aims to contest the claims that LLMs are necessarily incapable of acquiring meaning, and suggest that LLM meanings may already be similar to human-like meaning in many (but not all) ways. We argue that LLMs have likely already achieved some key aspects of meaning which, while imperfect, mirror the way meanings work in cognitive theories, as well as approaches to the philosophy of language.

LLMs are trained to predict words from massive datasets of text from the internet. They typically contain billions of parameters that are jointly optimized—at great computational, energy, and financial expense (Bender et al., 2021)—to make predictions about word occurrence from surrounding context. This setup differs from human language acquisition in data scale and format. On the other hand, the core objective of word prediction is a central piece of human language processing (e.g. Altmann and Kamide, 1999; Hale, 2001; Levy, 2008) and has long been shown capable of providing a learning signal from which linguistic structures and semantic categories can emerge (Elman, 1990).

A key debate in recent literature has been about *what* models trained in this manner come to know. A prominent view is that models trained only on text cannot acquire realistic meanings because they lack reference, or connection to objects in the real world. Bender and Koller (2020) illustrate this with an "octopus test." They imagine an octopus that learns to use words correctly by eavesdropping on a conversation between two people on land. If the octopus has no access to the referents of the words then there are gaps in its meaning for the words. For example, if the octopus must suddenly determine which *object* is a coconut, then its expertise in using the word "coconut" won't help. Its knowledge of co-occurrence statistics between "coconut" and other words won't help either since it also knows nothing about the referents of other words. The octopus simply does not have the required meanings to find the coconut. Humans learners don't have this problem because their input—and consequently representations—are tied to real-world referents. Bender & Koller's position is that no amount of predictive linguistic savvy can give models that are trained on text alone the knowledge of reference they need to acquire meanings.

Meaning and reference The octopus test assumes that reference determines meaning, but in fact cognitive scientists and philosophers have

found a variety of problems with this view. One is that there are many terms that are meaningful to us but have no discernible referent at all, such as abstract words like "justice" and "wit." People can think of new concepts like "aphid-sized accordion" that don't exist (thus no referent), or even terms that have no *possible* referent like "perpetual motion machine" or "imaginary cup of tea." We can think of concepts like "king of San Francisco" that pick out nobody, but are at least meaningful enough to reason about (for example: "If there was a King of San Francisco, he'd live in The Presidio."). Other examples show reference can be quite decoupled from meaning. Terms like "treaty" and "contract" are often thought to have a concrete referent, but what is important is actually an abstract entity: a treaty is still valid if the piece of paper is destroyed. Frege's example is of the "morning star" and the "evening star." Both are terms for the planet Venus, but were once conceived of as different entities without knowing that they are the same object.

This problem is not solely an issue with abstract concepts. Many concrete terms seem not to get their most important semantic properties from reference either. Consider an example like a "postage stamp." Everyone can conjure up an image of a typical, physical, postage stamp. But with further scrutiny, none of the concrete features of typical stamps seem necessary. We can imagine a country whose postage stamps were made of clear glass, for example, or stamps that were microscopic so that they could not be seen, or stamps that were paid for and tracked entirely online, or others that were larger than a house (for mailing very large packages). We could think of postage stamps that were *RFID* tags that went inside the envelope. Maybe intelligent ants would use postage stamps that were pheromones rather than paper; in the future we might have postage stamps that sprout wings and fly your letter to its intended location. Already we have stamps that only have barcodes and stamps that you can draw yourself. You know the term "postage stamp" but there is no way that you could have considered all of the possible referents yet, so reference cannot be what determines the concept.

Similar arguments were made by Wittgenstein (1953) when considering the concept of "water", leading him to conclude that reference plays little or no role in determining meaning in the general case. One way to explain these observations is to assume that our meaning for terms like "postage stamp" (or "water") may be primarily determined by the role these concepts play in some greater mental theory. Very roughly, people call something a "postage stamp" if you pay for it and then attach it to a letter in order to have the letter to be delivered. In this view, the meaning of the word is intrinsically intertwined with other concepts like "payment", "letter" and "delivery." The interrelation is key (see, e.g. Deacon, 1998; Santoro et al., 2021) because when the associated terms shift in meaning even slightly (e.g. credit cards were developed as wholly new a form of "payment") the meaning of "postage stamp" comes along for the ride (we know right away that they can be paid for by a credit card). Such relationships between concepts are *the* essential, defining, aspects of meaning and, in fact, possessing the appropriate relationships allows you to determine the reference. This view makes sense of the puzzling examples above.

Conceptual role theory In philosophy of mind, versions of this approach go by the name "conceptual role theory." Following Block (1998), consider statements in physics like $f = m \cdot a$. This equation is not exactly a definition of f (*force*), nor is it a definition of *mass* (m) or *acceleration* (a), but, is a statement about the interrelationship of f, m, and a. Most of us can't give much more detail about the ultimate physical reference of these terms—forces have something to do with interacting elementary particles (whatever those are), masses have something to do with the Higgs field (whatever that is)—but our thoughts about $f = m \cdot a$ certainly do not seem meaningless. Many believe that conceptual roles are one of the most promising ways to characterize human concepts *in general* (for an overview of this and competing theories, see, e.g., Margolis and Laurence, 1999). Murphy and Medin (1985) for example argue that our organization of categories is based on entire theories of structured conceptual domains (rather than simple features or similarities), an idea they trace to Quine (1977). Murphy and Medin note how we might reflexively consider a composite category such as "prime numbers or apples" to be an unnatural or incoherent set of entities. However, if we know someone called Wilma who is a number theorist grew up on an apple farm, then the category "topics of conversation with Wilma", involving the same constituent entities, seems perfectly reasonable. What is a natural category or concept depends on our mental conception of how the underlying pieces relate, and

concepts can even be assembled fluidly, in an ad hoc manner or context-dependent manner (Barsalou, 1983; Casasanto and Lupyan, 2015). Theories in cognition have also been explored in learning models (e.g. Goodman et al., 2011; Ullman et al., 2012) and experimentally shown to shape how children explore the world (e.g. Gopnik et al., 1999; Gopnik and Schulz, 2004; Bonawitz et al., 2012).

Conceptual role in LLMs If anything like this view is correct, then the search for meaning in learning models—or brains—should focus on understanding the way that the systems' internal representational states *relate to each other*. Once a learning model finds it probable that "postage stamps" are "affixed" to "letters" so they can be "delivered," then it has acquired some important pieces of conceptual role for these terms. It would not be possible to conclude anything about what meanings a system does and does not possess from its training data or architecture because these may not be informative about how the internal states relate to each other.

Relations between internal states have been long emphasized in cognitive theories (Shepard and Chipman, 1970; Deacon, 1998; Fodor and Pylyshyn, 1988), for example early attempts to discover the geometry of psychological space (Shepard, 1980) and more recent analyses of brain data based on representational similarity (Kriegeskorte et al., 2008). Elman (2004) argues for a closely related view of the mental lexicon in which a word's meaning is the effect it has on other mental states. In deep learning models, the relational geometry of vector representations have been examined for instance in analogy problems (Mikolov et al., 2013), match to human similarities (Hill et al., 2017), and encoding of humanlike gradient distinctions (Vulić et al., 2017). Grand et al. (2022) show that semantic embeddings from these models capture gradient scales of multiple features, like from "small" to "big" or "safe" to "dangerous." It is even possible to align the word representations acquired by text-based models across languages to translate between them effectively with no prior knowledge of which words or phrases should have the same meaning (Lample et al., 2017). Similarly, Abdou et al. (2021) show that a model trained on text can recover key geometry of color space, even without grounding; with a few examples of grounding, LLMs are able to align their structure with the real grounded one, suggesting that they already possess the right relations (Patel and Pavlick, 2021). Importantly, as the performance of LLMs has improved in recent years the extent to which their relational geometry reflects human data has also consistently increased (Peters et al., 2018; Devlin et al., 2018; Brown et al., 2020). Larger models also better reflect the human tendency for semantic or mental models to influence formal reasoning behaviour (Wason and Johnson-Laird, 1972), on challenging logic problems that are not observed in their training data (Dasgupta et al., 2022). Moreover, this increasing correspondence between LLMs and human data is not observed only behaviourally. Recent fMRI studies show that the semantic models that best account for the representational geometry and processing activity of human brains are precisely the neural network LLMs which are trained on the largest amount of data (Schrimpf et al., 2021; Goldstein et al., 2022; Kumar et al., 2022).

Many of the tasks that LLMs succeed on are ones that require maintaining the right relationships between concepts. Impressively, the largest models can now devise coherent narratives (Brown et al., 2020), extend stories (Xu et al., 2020; Li et al., 2021), answer factual questions (Jiang et al., 2021) solve Winograd Schema (Kocijan et al., 2020) and resolve complex quantitative reasoning problems (Lewkowycz et al., 2022). Increasingly, such models are even aware of the likelihood that they can answer a given question correctly; i.e. they have an explicit sense of the extent of their own knowledge (Kadavath et al., 2022). Each of these capacities requires, in some way or another, sensitivity to conceptual roles because the required words and concepts must be used jointly together in a coherent way that mimics how humans would.

Despite these empirical successes, there are many places where these models can still be improved (for detailed analyses, see, Lake and Murphy (2021); McClelland et al. (2020); Pavlick (2022)). Lake and Murphy (2021) emphasize the need for reference, inference, better and more robust compositionality, more structure and more consistent abstract reasoning. Models trained on multimodal datasets show better match to human judgements than those trained on text alone (Hill et al., 2016; De Deyne et al., 2021); at the same time, even multimodal LLMs are missing many aspects of a complete theory of semantics, including the ability to simulate situations in which their physical

or linguistic behaviour affects their environment (McClelland et al., 2020) as well as knowledge of the goals and desires that drive how people use words (Bisk et al., 2020; Lake and Murphy, 2021).

Our claim, then, is not that LLMs perfectly capture human concepts or perfectly reflect human meaning. Unlike the more radical perspectives entertained by Wittgenstein or Quine, we also do not consider that reference should play *no* role in a principled treatment of meaning. Instead, we find it productive to consider reference as just one (optional) aspect of a word's full conceptual role. It is relevant for some concepts (see Putnam, 1974) and not others—just like color, valence, or teleology is relevant for some concepts and not others. Experience of both agency and a perceptual environment similar to our own may lead to the richest, most human-like understanding of language in machines (Bisk et al., 2020; McClelland et al., 2020).

As these improvements are made, the models will come into closer alignment with humans, and each such improvement will enrich the model's sense of meaning. This process of progressive enrichment is also found in human concepts. When people discovered that H_2O was the chemical composition of water, they grew their conceptual network and even revised their reference for the term. There was no hard transition from a meaningless concept of "water" to a meaningful one. Some meaning was there all along because "water" had a conceptual role even before its chemical composition was known. What changed was the richness and interconnection of this concept—the way in which it was related to other concepts like "hydrogen" and "oxygen." In much the same way, we see no reason to assume that the world of a system that receives input from a single sensory modality is meaningless, even if the addition of further sensors provides clear enrichment. When thinking about improving LLMs it we should therefore consider ways to enrich the internal conceptual roles of these systems, including to better reflect the structure, inference and algorithmic sophistication of humans (Tenenbaum et al., 2011; Lake and Murphy, 2021; Rule et al., 2020).

Discovering conceptual role Conceptual role theory also provides a compelling way to understand learning, including the way in which neural networks may come to represent symbolic processes. A symbol like AND only means logical conjunction if it interacts (composes) with others symbols like $TRUE$ and $FALSE$ in the appropriate way—i.e. when it has the right conceptual role in the broader system of symbols. The technique of *church-encoding* in mathematical logic (see Pierce, 2002) provides a way to understand how such roles may be learned within neural networks or dynamical systems (Piantadosi, 2021). In church-encoding, a representation is constructed in one system (e.g. lambda calculus or a neural network) in order to mimic the behavior of another system (e.g. boolean logic) in the sense that the representations in the first system interact with each other in a way that yields the desired conceptual roles of the second. Piantadosi (2021) shows how a church-encoding learner could acquire structures like logic, lists, trees, hierarchies, numbers, quantifiers, and recursion without possessing them to start, and how this metaphor may provide an "assembly language" that translates from symbolic computational or cognitive theories into underlying implementations. In this view, neural networks would train their parameters so that their internal, intrinsic dynamics church-encode the conceptual roles of a targeted domain.

The key question for LLMs is whether training to predict text could actually support discovery of conceptual roles. To us the question is empirical, and we believe has been answered in a promising, partial affirmative by studies showing success on tasks that require knowledge of relationships between concepts. Text provides such clues to conceptual role because human conceptual roles generated the text. Analogously, it is possible to build a theory of gravity from measurements of the moon's movement because gravity *generated* these movements; the goal of essentially all inductive learning techniques is to invert from observations to likely generating processes or parameters. Moreover, the entailment relationships between sentences are often intrinsically related to analogous patterns in thought (e.g. Fodor and Pylyshyn, 1988), meaning that a model which captures how sentences relate to each other might indeed capture how thoughts relate to each other. One helpful analogy is that of *embedding theorems* in dynamical systems (e.g. Packard et al., 1980; Takens, 1981) which allow some properties of systems to be recovered from seemingly impoverished representations of their state. In the paper "Geometry from a Time Series", Packard et al. (1980) show, remarkably, that one can sometimes reconstruct the geometry of a

multi-dimensional dynamical system from a *one*-dimensional projection of its state. Thus, information about high-dimensional state (sometimes essentially all of it) can be decoded from the trajectory of low-dimensional projection. People use concepts in thinking and reasoning based on their meaning, and text is a low-dimensional projection of some of these patterns of use, so it is plausible that some properties of the real meaning could be inferred from text. At the very least, embedding theorems illustrate that there may not be a simple way to intuit what a learner can or cannot deduce about the underlying mental states from text alone.

The protein folding neural network AlphaFold (Jumper et al., 2021) provides further evidence that transformer-based networks can infer and generalise complex latent multi-entity structures. AlphaFold is trained to predict the configuration of single proteins only, but acquires actionable knowledge of concepts not explicitly present in its training data. Despite never seeing a zinc ion, AlphaFold often perfectly infers its location and places all the protein side chains correctly right around it. Some proteins only fold with multiple copies of themselves (homomers). Again, AlphaFold has never seen more than one copy of a protein, but often infers both the number of required copies and their relative placement correctly. This suggests that the process of predicting the structure of single proteins enabled the AlphaFold network to infer non-trivial facts about chemistry and biology.

Communicative intentions Separable from meaning and reference, many have also rejected the idea that LLMs produce language with *intention* (Bender et al., 2021; Bender and Koller, 2020). Because LLMs are trained only on sequence prediction, they are argued to be, "stochastic parrots" (Bender et al., 2021) or just sophisticated "auto-complete" algorithms[1] that cannot access the intentions of those who produced their training data, and themselves produce language without intending anything in particular. But in our view, a key difference between autocompleting parrots and LLMs is that the latter have rich, causal, and structured internal states.

One view of intent is semantic, corresponding to whether the language they produce arises from an internal representation of (intended) meaning. The conversion of internal states into language and back *is* the essential function of LLMs, embedded in their architecture and training. We have argued that LLM's internal state has some notions of conceptual role, so LLM's utterances have the semantic intent corresponding to these roles.

Another view of intent is pragmatic, asking what might be achieved by producing a sentence. This corresponds to asking whether they engage in any goal-directed *planning* (Russell, 2010) when producing language. We consider it probable that multi-layer LLMs do, in an emergent, implicit sense, execute a form of planning as part of the process of repeated (self-attention-based) analyses of current and past inputs. These computations likely involve representation of the current situation and at least implicit evaluation of consequences of utterances. Recent work has argued that LLMs posses model-like belief structures (Hase et al., 2021) and update representations of dynamic semantics, objects and situations, throughout a discourse (Li et al., 2021). These emergent semantics causally determine LLM output. Of course, differences between humans and LLMs in their training experience and objectives mean that the planning process in LLMs is less explicit and sophisticated than in humans (making errors more likely, for instance, in cases of hypothetical or counterfactual reasoning (Ortega et al., 2021)).

Conclusion Bender & Koller argue that text-based LLMs will never have meaning because these models lack reference. However, they do not demonstrate that reference is the key to meaning—instead they assume it. As we have argued, this assumption is hard to reconcile with theories of cognition and the phenomena that motivate them. People are happy to think about concepts without referents and otherwise often don't know many details of reference. Meaning instead seems to come from the way concepts relate to *each other*. It is these interrelations that LLMs know something about since their internal geometries and trajectories approximate those of humans. Like people who don't know that water is H_2O and so could not pick it out based on chemical composition, Bender & Koller's octopus lacks some aspects of conceptual role like physical appearance. But, both the octopus and people know other parts of conceptual role that are sophisticated in their own right. If theories about conceptual role are the correct account, then LLMs likely already share the foundation of how our own concepts get their meaning.

[1] https://garymarcus.substack.com/p/nonsense-on-stilts

Acknowledgements

We are grateful to Ev Fedorenko, Adam Santoro, MH Tessler, Dileep George, John Hale, Miguel Figurnov, John Jumper, Dharsh Kumaran, Matt Botvinick, Paul Smolensky and Chris Dyer for detailed comments.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? A case study in color. *arXiv preprint arXiv:2109.06129*.

Gerry TM Altmann and Yuki Kamide. 1999. Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition*, 73(3):247–264.

Lawrence W Barsalou. 1983. Ad hoc categories. *Memory & cognition*, 11(3):211–227.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Emily M. Bender and Alexander Koller. 2020. 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*, pages 5185–5198, Online. Association for Computational Linguistics.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. 2020. Experience grounds language. *arXiv preprint arXiv:2004.10151*.

Ned Block. 1998. Semantics, conceptual role. *Routledge encyclopedia of philosophy*, 8:652–657.

Elizabeth Baraff Bonawitz, Tessa JP van Schijndel, Daniel Friel, and Laura Schulz. 2012. Children balance theories and evidence in exploration, explanation, and learning. *Cognitive psychology*, 64(4):215–234.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Daniel Casasanto and Gary Lupyan. 2015. All Concepts Are Ad Hoc Concepts. In E. Margolis and S. Laurence, editors, *The Conceptual Mind: New directions in the study of concepts*, pages 543–566. Cambridge: MIT Press.

Ishita Dasgupta, Andrew K Lampinen, Stephanie CY Chan, Antonia Creswell, Dharshan Kumaran, James L McClelland, and Felix Hill. 2022. Language models show human-like content effects on reasoning. *arXiv preprint arXiv:2207.07051*.

Simon De Deyne, Danielle J Navarro, Guillem Collell, and Andrew Perfors. 2021. Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, 45(1):e12922.

Terrence William Deacon. 1998. *The symbolic species: The co-evolution of language and the brain*. 202. WW Norton & Company.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Jeffrey L Elman. 2004. An alternative view of the mental lexicon. *Trends in cognitive sciences*, 8(7):301–306.

Jerry A Fodor and Zenon W Pylyshyn. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. 2022. Shared computational principles for language processing in humans and deep language models. *Nature neuroscience*, 25(3):369–380.

Noah D Goodman, Tomer D Ullman, and Joshua B Tenenbaum. 2011. Learning a theory of causality. *Psychological review*, 118(1):110.

Alison Gopnik, Andrew N Meltzoff, and Patricia K Kuhl. 1999. *The scientist in the crib: Minds, brains, and how children learn.* William Morrow & Co.

Alison Gopnik and Laura Schulz. 2004. Mechanisms of theory formation in young children. *Trends in cognitive sciences*, 8(8):371–377.

Gabriel Grand, Idan Asher Blank, Francisco Pereira, and Evelina Fedorenko. 2022. Semantic projection recovers rich human knowledge of multiple object features from word embeddings. *Nature Human Behaviour*, pages 1–13.

John Hale. 2001. A probabilistic earley parser as a psycholinguistic model. In *Second meeting of the north american chapter of the association for computational linguistics*.

Peter Hase, Mona Diab, Asli Celikyilmaz, Xian Li, Zornitsa Kozareva, Veselin Stoyanov, Mohit Bansal, and Srinivasan Iyer. 2021. Do language models have beliefs? methods for detecting, updating, and visualizing model beliefs. *arXiv preprint arXiv:2111.13654*.

Felix Hill, Kyunghyun Cho, Sébastien Jean, and Yoshua Bengio. 2017. The representational geometry of word meanings acquired by neural machine translation models. *Machine Translation*, 31(1):3–18.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. 2021. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know.

Vid Kocijan, Thomas Lukasiewicz, Ernest Davis, Gary Marcus, and Leora Morgenstern. 2020. A review of winograd schema challenge datasets and approaches. *arXiv preprint arXiv:2004.13831*.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A Bandettini. 2008. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, page 4.

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. 2022. Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *bioRxiv*.

Brenden M Lake and Gregory L Murphy. 2021. Word meaning in minds and machines. *Psychological review*.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. 2022. Solving quantitative reasoning problems with language models. *arXiv preprint arXiv:2206.14858*.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. *arXiv preprint arXiv:2106.00737*.

Eric Ed Margolis and Stephen Ed Laurence. 1999. *Concepts: Core Readings.* The MIT Press.

James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2020. Placing language in an integrated understanding system: Next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, 117(42):25966–25974.

Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751.

Gregory L Murphy and Douglas L Medin. 1985. The role of theories in conceptual coherence. *Psychological review*, 92(3):289.

Pedro A Ortega, Markus Kunesch, Grégoire Delétang, Tim Genewein, Jordi Grau-Moya, Joel Veness, Jonas Buchli, Jonas Degrave, Bilal Piot, Julien Perolat, et al. 2021. Shaking the foundations: delusions in sequence models for interaction and control. *arXiv preprint arXiv:2110.10819*.

Norman H Packard, James P Crutchfield, J Doyne Farmer, and Robert S Shaw. 1980. Geometry from a time series. *Physical review letters*, 45(9):712.

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International Conference on Learning Representations*.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8:447–471.

ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. 2018. Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12.

Steven T Piantadosi. 2021. The computational origin of representation. *Minds and machines*, 31(1):1–58.

Benjamin C Pierce. 2002. *Types and programming languages*. MIT press.

Hilary Putnam. 1974. Meaning and reference. *The journal of philosophy*, 70(19):699–711.

W.V.O. Quine. 1977. Natural kinds. In S.P. Schwartz, editor, *Naming, necessity, and natural kinds*, pages 155–175. Ithaca, NY: Cornell University Press.

Joshua S Rule, Joshua B Tenenbaum, and Steven T Piantadosi. 2020. The child as hacker. *Trends in cognitive sciences*, 24(11):900–915.

Stuart J Russell. 2010. *Artificial intelligence a modern approach*. Pearson Education, Inc.

Adam Santoro, Andrew Lampinen, Kory Mathewson, Timothy Lillicrap, and David Raposo. 2021. Symbolic behaviour in artificial intelligence. *arXiv preprint arXiv:2102.03406*.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Roger N Shepard. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398.

Roger N Shepard and Susan Chipman. 1970. Second-order isomorphism of internal representations: Shapes of states. *Cognitive psychology*, 1(1):1–17.

Floris Takens. 1981. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer.

Joshua B Tenenbaum, Charles Kemp, Thomas L Griffiths, and Noah D Goodman. 2011. How to grow a mind: Statistics, structure, and abstraction. *science*, 331(6022):1279–1285.

Tomer D Ullman, Noah D Goodman, and Joshua B Tenenbaum. 2012. Theory learning as stochastic search in the language of thought. *Cognitive Development*, 27(4):455–480.

Ivan Vulić, Daniela Gerz, Douwe Kiela, Felix Hill, and Anna Korhonen. 2017. Hyperlex: A large-scale evaluation of graded lexical entailment. *Computational Linguistics*, 43(4):781–835.

Peter Cathcart Wason and Philip Nicholas Johnson-Laird. 1972. *Psychology of reasoning: Structure and content*, volume 86. Harvard University Press.

Ludwig Wittgenstein. 1953. *Philosophical investigations*. John Wiley & Sons.

Peng Xu, Mostofa Patwary, Mohammad Shoeybi, Raul Puri, Pascale Fung, Anima Anandkumar, and Bryan Catanzaro. 2020. Megatron-cntrl: Controllable story generation with external knowledge using large-scale language models. *arXiv preprint arXiv:2010.00840*.