The Vector Grounding Problem

Dimitri Coelho Mollo
Department of Historical, Philosophical and Religious Studies
Umeå University
dimitri.mollo@umu.se

Raphaël Millière
Philosophy Department
Macquarie University
raphael.milliere@mq.edu.au

Abstract

The remarkable performance of large language models (LLMs) on complex linguistic tasks has sparked debate about their capabilities. Unlike humans, these models learn language solely from textual data without directly interacting with the world. Yet they generate seemingly meaningful text on diverse topics. This achievement has renewed interest in the classical 'Symbol Grounding Problem' – the question of whether the internal representations and outputs of symbolic AI systems can possess intrinsic meaning that is not parasitic on external interpretation. Although modern LLMs compute over vectors rather than symbols, an analogous problem arises for these systems, which we call the Vector Grounding Problem. This paper has two main goals. First, we distinguish five main notions of grounding that are often conflated in the literature, and argue that only one of them, which we call referential grounding, is relevant to the Vector Grounding Problem. Second, drawing on philosophical theories of representational content, we provide two arguments for the claim that LLMs and related systems can achieve referential grounding: (1) through preference fine-tuning methods that explicitly establish world-involving functions, and (2) through pre-training alone, which in limited domains may select for internal states with world-involving content, as mechanistic interpretability research suggests. Through these pathways, LLMs can establish connections to the world sufficient for intrinsic meaning. One potentially surprising implication of our discussion is that that multimodality and embodiment are neither necessary nor sufficient to overcome the Grounding Problem.

1. Introduction

The remarkable performance of Large Language Models (LLMs) in sophisticated linguistic and cognitive tasks – once considered reliable indicators of human-like intelligence – has rekindled philosophical discussions on the nature of linguistic competence (Millière & Buckner 2024, Millière forthcoming). Despite being trained exclusively on text-based data without direct interaction with the physical world, these models exhibit the ability to generate coherent and contextually relevant paragraphs, answer questions, and solve commonsense reasoning tasks in a wide range of domains.

This ability raises intriguing questions about the relationship between language, meaning, and reference in disembodied artificial systems that learn from input of raw text.[1]

In this paper, we revisit the classical 'Symbol Grounding Problem', originally formulated for symbolic AI systems, in light of contemporary Generative Pre-Trained Models, and especially LLMs. The core issue can be summarized as follows: if AI systems are designed to process linguistic inputs without direct interaction with the world, can their internal representations and outputs possess any meaning beyond the interpretations that we, as intelligent beings embedded in the world, project onto them? We use the term *intrinsic meaning* to capture this notion of interpretation-independent meaning, distinguishing it from meaning that is imposed from outside the system itself, which is thus extrinsic.[2] We argue that a similar conundrum arises for Generative Pre-Trained Models, which we dub the *Vector Grounding Problem*. Since these systems compute over continuous vectors rather than discrete symbols, the problem assumes a slightly different form. After outlining this new version of the classical problem, we suggest that there are strong reasons to believe that a basic form of grounding is possible in current LLMs.

Specifically, we contend that the learning process of LLMs establishes functional roles for their internal states that go beyond mere linguistic prediction. Although LLMs learn from text-based data, they undergo selection pressures that can favour internal states tracking features of the world whenever extra-linguistic criteria influence the optimization objective. This can occur directly by fine-tuning LLMs on objectives that reflect extra-linguistic norms – such as factuality – or indirectly via the implicit extra-linguistic success conditions of latent tasks embedded within the training data or provided in context. When internal states that were selected to respond differentially to specific worldly conditions causally influence the generation of linguistic outputs, those outputs become grounded in the world in a way that makes them meaningful independently of human interpretation. The crucial question is thus not whether outputs can be interpreted as meaningful by humans (they obviously can), but whether the outputs themselves stand in the right kinds of relations to the world such that they would be meaningful even in the absence of external interpretation.

Importantly, this paper does not address whether LLMs and related models *understand* language (or anything at all), engage in linguistic acts, or possess cognitive or mental properties.[3] We do not claim that these models 'know' what they mean, or even what meaning is, nor that they have any conception of truth or falsity. Our sole contention is that the problem of grounding internal states and outputs in the world is solvable in principle, and in some cases arguably already in practice, for LLMs and related models. While having grounded internal states may be a crucial step toward understanding, agency, knowledge, and various other cognitive and mental capacities, it is unlikely that they are, by themselves, sufficient for those capacities. This clarification is especially important, as the current debate occasionally blurs these distinctions – a conflation we strive to avoid in this paper.

The paper proceeds as follows. First, we introduce the Symbol Grounding Problem in the context of classical AI (section 2). Next, we present the Vector Grounding Problem as it applies to LLMs and related models (section 3). We then differentiate between various senses of grounding, showing that

[1] For recent discussion, see Butlin (2021), Cappelen & Dever (2021), Pavlick (2022), Chalmers (2023), Mandelkern & Linzen (2024), Lederman & Mahowald (2024), Miracchi Titus (2024), Grindrod (2024), Borg (2025), Pepp (2025).

[2] This follows Harnad (1990)'s terminology. Intrinsic meaning, as we understand it here, is fully compatible with externalist views about how states and outputs come to acquire interpretation-independent meaning and reference. The intrinsic/extrinsic distinction cuts across the internalism/externalism one. We thank an anonymous referee for prompting this clarification.

[3] See Goldstein & Levinstein (2024) and Chalmers (2025) for discussions that also draw on theories of content but consider whether LLMs can be ascribed propositional attitudes, and Stoljar & Zhang (2025) for a discussion of whether LLMs can think.

only one of them – *referential grounding* – is essential to the Vector Grounding Problem (section 4). Drawing on philosophical theories of representational content, we articulate a set of conditions that must be fulfilled by a system to achieve referential grounding (section 5). Finally, we argue that at least some LLMs and related multimodal models satisfy these requirements, and therefore have internal representations and outputs with intrinsic meaning (section 6).

2. The Classical Symbol Grounding Problem

Research on artificial intelligence has historically followed two competing paradigms: classical (or symbolic), and connectionist. The classical paradigm focuses on systems that represent and manipulate information using discrete symbols, based on explicit rules or procedures. Symbolic manipulation is purely syntactic; it is based on the shapes of symbols and the way in which they can be combined into complex symbolic structures (Simon & Newell 1971). In contrast, the connectionist paradigm focuses on artificial neural networks that process information in a distributed way through interconnected nodes, learning patterns from data rather than following explicit symbolic rules (Mcclelland et al. 1986).

The Symbol Grounding Problem, as identified by Harnad (1990), asks how the meaning of symbols in a formal symbol system can be intrinsic to the system itself, rather than merely parasitic on the meanings in the minds of external interpreters. More precisely:

> **Symbol Grounding Problem.** How can symbol tokens, manipulated solely on the basis of their shapes according to syntactic rules, acquire meaning and refer to entities and properties in the real world – rather than merely relating other meaningless symbol tokens – without depending on the semantic interpretations provided by external meaning-bearing systems?[4]

Harnad (1990) provides a simple and powerful intuition pump to motivate the existence of a Symbol Grounding Problem in classical, symbolic AI: the Chinese Dictionary. Take a system, be it biological or artificial, which knows no natural language. Suppose that we want it to learn Chinese (or any other language), but the only resource made available to the system is a single-language dictionary, say a comprehensive Chinese-Chinese dictionary. Will the system be able to learn Chinese just by using the dictionary? The thought experiment invites the intuition that the system would not be able to do so. After all, the most it can do is to look up strings of symbols and see what other strings of symbols follow them. The former would be the words to be defined, the latter their definitions. However, by themselves the symbols are just meaningless shapes followed by other shapes, and it seems intuitively hopeless that the meaning of words (parts of those strings of symbols) could be learned this way. What seems to be missing is a way out from the symbolic 'merry-go-round': to connect those symbols in the dictionary to what they pick out in the world. But how can that be possible if the system has no access to the world, but only to further symbols?

Symbolic AI systems, according to Harnad, find themselves in this predicament. All they have access to are strings of symbols and how they follow each other. There seems to be no way for connections between such symbols and the world to be established, and thus no way for the symbols to have intrinsic meaning. Humans, who do have representations arguably acquired by means of engagement with the world, are, on the other hand, able to interpret those symbolic representations and outputs as meaningful. The meaning of the representations is hence extrinsic, depending on human interpreters. Take out the interpreter, and no meaning is to be found.

[4] It is common to take Harnad's Symbol Grounding Problem as an elaboration of an earlier, highly influential criticism of AI: Searle (1980)'s Chinese Room thought experiment. Harnad (1990) himself followed this line.

In brief, symbolic AI systems, the argument goes, manipulate symbols that are not connected to the world, that are not grounded in the world; and thereby have no intrinsic meaning. Classical AI systems, in other words, suffer from the Symbol Grounding Problem.

3. The Vector Grounding Problem

Connectionist models have come a long way since the 1980s, with deep neural networks (DNNs) emerging as powerful tools for various tasks (LeCun et al. 2015). Recent research has converged onto training deep neural networks on vast amounts of data to perform a broad range of tasks they were not explicitly designed for. Such DNNs are routinely called 'Foundation Models' (Bommasani et al. 2022), though we will use the more neutral label 'Generative Pre-Trained Models'. Many among the most popular Generative Pre-Trained Models today are powered by the Transformer architecture, first introduced by Vaswani et al. (2017). Transformers use an attention mechanism to model global dependencies between sequential elements in the input, such as sentences in natural language. The general inductive biases supplied by this mechanism make Transformers suitable for many domains, including natural language processing and computer vision.

Some Generative Pre-Trained Models, called *Large Language Models* (LLMs), are trained exclusively on text. State-of-the-art LLMs like GPT-4 and Claude have over 10^{11} parameters and are trained on massive corpora spanning hundred of billions of words, containing a significant subset of all written text on the internet. After training, LLMs can generate well-written and coherent paragraphs about virtually any topic. Furthermore, given explicit instructions in natural language in the input sequence, or 'prompts', they can perform a broad range of tasks, including: summarising long-form content, translating between multiple languages, answering complex questions, engaging in commonsense reasoning, solving simple logic and maths problems, generating code for simple programs, or explaining jokes (Srivastava et al. 2023).

LLMs process input text into a sequence of tokens (i.e., full words, word-parts such as '-ing', punctuation marks, and other linguistic symbols). LLMs encode each token as a vector of real-valued numbers – known as an *embedding* – in a high-dimensional vector space, and estimate the probabilities of all possible next tokens in the sequence based on the previous embeddings. The resulting prediction of the next token in the sequence is then compared to the actual next token, and the model's parameters are subsequently updated to make the correct token more probable in that context, and other tokens less probable. While LLMs do not treat words as the main units of processing, the combination of tokens can form an indefinite number of possible words, and we can view the vectors for such composing tokens as jointly forming a word embedding. This training process (confusingly known as pre-training) encourages the model to produce context-sensitive embeddings, with semantically similar words having vectors that are close together in the high-dimensional space.

There is a deep connection between the way in which LLMs learn embeddings from linguistic corpora, and the old idea from structural linguistics according to which word meaning can be inferred from the contexts in which they appear, known as the 'distributional hypothesis' (Harris 1954, Firth 1957). In LLMs, semantic similarity is determined based on distribution across contexts: words that appear in similar contexts are modelled by nearby vectors. The dimensions of the vectors are tuned by the model to best predict surrounding words, and capture distributional patterns. Although individual dimensions are not typically interpretable, the positions of vectors in the space are a result of statistical patterns in the linguistic data, and reflect semantic similarities in light of how the words are used

in context.[5] In this way, embeddings in LLMs are not arbitrary or conventional – as the symbols in classical AI systems typically are – since they are shaped by statistical patterns in language.

However, LLMs appear to face a similar predicament to that of traditional symbolic AI systems when it comes to grounding and intrinsic meaning. Their reliance on continuous vectors, rather than discrete symbols, seems to make no difference to escaping the merry-go-round of representations, and being connected to the external world – LLMs are trapped in a carousel of numbers, rather than symbols. Much like their classical counterparts, LLMs manipulate vectors that do not seem grounded in the world and thus lack intrinsic meaning. Consequently, they grapple with an analogous difficulty, which we call the *Vector Grounding Problem*.

Mirroring Harnad's 1990 Symbol Grounding Problem, the Vector Grounding Problem asks how the vector embeddings inside Generative Pre-Trained Models can have intrinsic meanings, rather than merely being parasitic on the meanings in the minds of external interpreters. More precisely:

> **Vector Grounding Problem.** How can vector embeddings, manipulated solely based on their numerical values according to algebraic operations, acquire meaning and refer to entities and properties in the real world – rather than merely relating other meaningless vectors – without depending on the semantic interpretations provided by external meaning-bearing systems?

Both Grounding Problems, as stated here, focus on internal states, such as symbols and vector embeddings. If the Grounding Problems are solved, and such internal states come to have intrinsic meaning, we get straightforward reasons to think that the systems in question produce outputs that are intrinsically meaningful. For the meaning of the outputs would thus derive from the meaning of the system's internal representations, rather than from the interpretations given by human users.[6]

It is worth noting that the Grounding Problems are particularly relevant to systems that generate novel outputs. This sets them apart from systems that merely reproduce outputs generated by humans, such as paper copiers. When a human-generated text is copied, there is no question about its intrinsic meaning: it was produced by a system with meaningful internal representations to convey a specific meaning to the reader. In contrast, generative systems like LLMs produce outputs that are not, for the most part, mere copies of human-generated text; they generate entirely novel sentences whose meanings cannot be equated to what any specific human has ever thought or said. Unlike with paper copiers, the question of whether LLM outputs are meaningful cannot be straightforwardly explained by appealing to a human author.[7]

Bender & Koller (2020) propose a thought-provoking thought experiment, the "Octopus Test", to pump intuitions about the inescapability of the Vector Grounding Problem for LLMs:

> **The Octopus Test.** Two English-speaking castaways find themselves stranded on neighbouring islands, separated by treacherous waters. Fortuitously, they discover telegraphs left by previous inhabitants, connected via an underwater cable, which enables them to communicate by exchanging telegraphic messages. Unbeknownst to them, a superintelligent octopus inhabits these waters and taps into the underwater cable, intercepting their messages. Though the octopus lacks any knowledge of English, its superintelligence allows it to detect statistical patterns in the telegraphic messages and to form an accurate representation of the statistical relationships between various

[5] Vision-Language Models (VLM) that can process images in addition to text work in a similar way: they encode text and images as vectors in a single high-dimensional space. We will discuss whether these models have specific implications for the grounding problem in subsection 7.3.

[6] In such a case, human users could misinterpret the system's output if their interpretation does not conform to what the system actually represented – in a way akin to how humans can misinterpret each other.

[7] We thank an anonymous referee for pressing us on this point.

> telegraphic signals. The octopus decides to sever the underwater cable and position itself at the ends of the two resulting cable segments, receiving and replying to telegraphic signals from both castaways based on the statistical patterns it has identified. Whether or not the castaways notice this change, the messages sent by the octopus intuitively seem to hold no intrinsic meaning. After all, the octopus simply adheres to the statistical patterns it has learned from listening in on the previous exchanges between the humans, without any understanding of the human interpretation of the signals, such as 'coconut' or 'tree'. Furthermore, the octopus likely does not comprehend that the signals possess meaning or serve a communicative function.[8]

As should be clear, the octopus in this thought experiment is operating roughly as an extremely powerful LLM would. Regardless of how the octopus represents the signals and their statistical relations, be it through vectors or other means, it only has access to the signals and their statistical relations – without any connection to the entities and properties in the real world that the words encoded in those signals represent. Consequently, even if the octopus can produce outputs that appear sensible to humans in all situations, these outputs lack intrinsic meaning.[9] The Octopus Test implies that LLMs generate outputs as devoid of intrinsic meaning as those produced by the superintelligent octopus.[10] It is humans, whether castaways or otherwise, who interpret these outputs as meaningful because they resemble meaningful words and sentences. LLMs have no access to coconuts, trees, or anything else beyond language. The vectors they receive and manipulate during internal processing are merely meaningless arrays of numbers, ungrounded in anything outside statistical patterns in language, just like the outputs they generate. In other words, the Octopus Test invites the intuition that all meaning in the outputs produced by LLMs is extrinsic, i.e., it lies in the eye of the beholder, their human interpreters. This would be so, we might add, because LLMs are susceptible to the Vector Grounding Problem.

Throughout the remainder of this paper, we will challenge this claim and counter the intuition primed by the Octopus Test. We aim to show that, contrary to first appearances, LLMs indeed possess the features required to generate meaningful outputs based on meaningful internal representations.

Before developing this argument further, it is essential to clarify the target notion of grounding.

4. Five notions of Grounding

Various notions of grounding have been invoked by cognitive scientists, philosophers, and AI researchers as central for endowing internal representations and outputs with meaning, occasionally leading to confusion. In this section, we endeavour to bring clarity to this debate by identifying five primary notions of grounding at play in discussions about intrinsic meaning in cognitive systems, both biological and artificial (Figure 1). Furthermore, we contend that only one of these five notions of grounding is crucial for addressing the Vector Grounding Problem. If LLMs possess features that

[8] Bender & Koller (2020) emphasise the fact that the octopus lacks communicative intent when generating telegraphic signals. Since our focus here is the narrower, more basic question of vector grounding, rather than the more complex capacities related to language comprehension and communication, we will set those considerations aside.

[9] Our version of the Octopus Test goes beyond Bender & Koller (2020) by assuming that it is possible the humans would never notice they were no longer communicating with each other. In contrast, Bender & Koller (2020) contend that the octopus would fail to produce appropriate outputs to questions involving concrete objects it has never perceived, such as 'how do I build a coconut catapult?'. This claim seems to be challenged by the often plausible replies that current LLMs provide to this sort of question. We should expect the superintelligent octopus to generate even more reliably plausible responses.

[10] Arguably, LLMs' outputs would be even more devoid of intrinsic meaning than the octopus', given that the latter is a biological organism with cognitive states, motivation, superintelligence, and other features that LLMs lack. Similar considerations apply to the comparison between LLMs and parrots (Bender et al. 2021). In the case of these animals, even if their outputs carry some meaning, they typically do not align with the meaning of the language they produce.

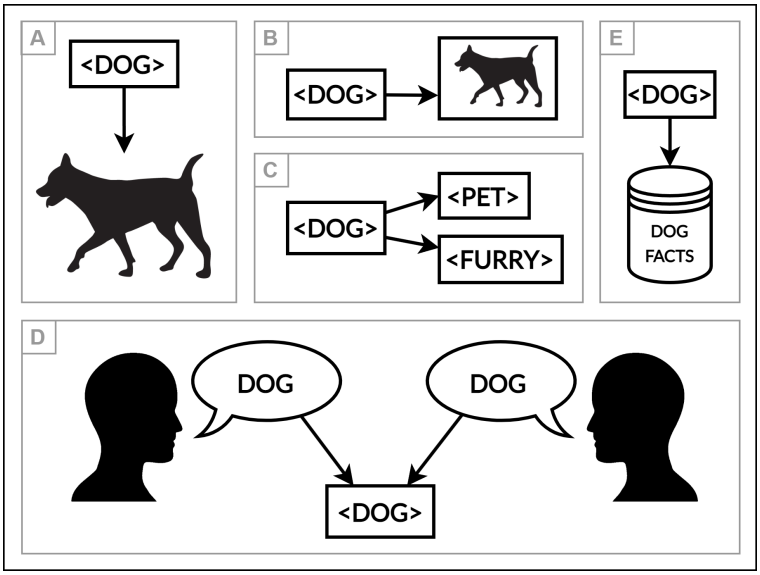

Figure 1 | Five notions of grounding. A. *Referential grounding*: a lexical representation (<DOG>) is connected to its worldly referent (a dog). B. *Sensorimotor grounding*: a lexical representation (<DOG>) is connected to a sensory representation (image of a dog). C. *Relational grounding*: a lexical representation (<DOG>) is connected to other lexical representations (<PET>, <FURRY>). D. *Communicative grounding*: Two speakers calibrate their interpretation of an exchange to make use of the same lexical concept (<DOG>). E. *Epistemic grounding*: A lexical representation (<DOG>) is connected to information stored in a knowledge base (about dog facts).

enable them to satisfy the criteria for this particular type of grounding, they become candidates for having representations and outputs with intrinsic meaning. With this in mind, let us examine each notion of grounding in turn.

4.1. Referential Grounding

The first notion of grounding, which we call *referential grounding*, closely aligns with the philosophical notion of reference.[11] It is the type of grounding that connects representations to things in the world. Reference can be loosely described as a relation that enables representations to 'hook onto' worldly entities or properties.[12] For example, the word 'Paris' refers to the city of Paris, and arguably, so does a map of Paris. Both representations allow us to receive and convey information about the actual city of Paris. Similarly, many of our thoughts seem to hook onto specific entities and properties in the world, allowing us to entertain thoughts that refer to Paris, such as beliefs about it being the capital of France or desires to visit the city.

Referential grounding, as we use the term, is whatever makes it so that representations, whether public (e.g., words and maps) or internal (e.g., beliefs and subpersonal states), hook onto the world. It is plausible that different kinds of representation require different means for hooking onto the world

[11] Partee (1981) uses the phrase "language-to-world grounding" to designate a closely related notion, which she contrasts with "language-to-language grounding" (see subsection 4.3 below for discussion of the latter). While Partee's notion of language-to-world grounding focuses on what ties down the intensions of linguistic expressions with respect to their extra-linguistic content, our notion of referential grounding is broader, including the relation between internal representations and their worldly referents.

[12] Much philosophical work has been dedicated to examining the nature and varieties of reference. For our limited purposes, however, the intuitive notion captured by the 'hooking' metaphor will suffice.

in the relevant sense, that is, to be referentially grounded in specific worldly entities and properties (Coelho Mollo 2015).

4.2. Sensorimotor Grounding

Another type of grounding that is often appealed to, especially in cognitive psychology, is what we call *sensorimotor grounding*. For an internal representation to be grounded in this sense is for it to be fundamentally linked to sensorimotor representations available within a cognitive system. For instance, the lexical representation 'football' is sensorimotorically grounded if its representation in a cognitive system can be reduced to, or is intrinsically connected to, sensorimotor representations of kicking movements, of the visual perception of a football pitch, of the auditory perception of the sound of a ball being kicked, etc. This kind of grounding plays a prominent role in theories in cognitive psychology that embrace a moderate form of embodied cognition, such as those positing the existence of "perceptual symbol systems" in human cognition (Barsalou 1999, Pulvermüller 1999).

Interestingly, sensorimotor grounding is also the solution proposed by Harnad (1990) to his Symbol Grounding Problem. He argued that the symbols in classical AI systems needed to be grounded in two types of representation: (a) iconic representations, which depict transformations of the projections of object 'shapes' on a system's sensory surfaces; and (b) categorical representations, which represent invariant sensory features that facilitate the identification and discrimination of objects in different categories. Harnad hypothesised that connectionist approaches to AI would be better equipped to provide the necessary iconic and categorical representations, which could then complement symbolic AI systems and result in hybrid symbolic-connectionist systems.

We will come back to Harnad's suggestion in subsection 4.6, arguing that it fails to solve the Symbol Grounding Problem, and does not help with the Vector Grounding Problem either.

4.3. Relational Grounding

A third notion of grounding, that we may call *relational grounding*, is occasionally used to refer specifically to intra-linguistic relationships, in accordance with the idea that a word's meaning in a language is partly determined by its relations to other words.[13] Partee (1981) introduces the notion of "language-to-language grounding", referring to the way in which a language learner acquires knowledge about the meanings of words by learning semantic connections between words in the language. According to Partee, one key way in which this happens is through accepting certain sentences as true, particularly those that are regarded as impervious to empirical challenge within the linguistic community. For example, learning that "bachelors are unmarried males" helps ground the meaning of 'bachelor' through its relation to 'unmarried' and 'male'.

Relational grounding is closely related to inferential role semantics (Block 1986, Brandom 1994). Inferential role semantics identifies the meaning of a linguistic expression with its inferential relations to other expressions. For example, the word 'and' plays a logical role in reasoning that determines when compound statements are true or false based on the truth or falsity of their components. Likewise, words like 'larger' or 'slower' have inferential roles in comparisons, allowing us to determine comparisons between items or the relative properties of a single item. Knowledge of such relations between words underlies our ability to draw material inferences in language and thought.

[13] The label is somewhat under-descriptive, since all forms of grounding are relational, though in different ways. What we wish to underline, however, is that on this sense of grounding the meaning of representations depends on the relations they stand in to other representations in the same representational system.

Finally, there is a clear connection between Partee's "language-to-language grounding" and distributional semantics. Distributional semantics seeks to analyse word meanings in terms of word-to-word relationships. However, instead of drawing on linguistic analyses of word meaning, it relies on analysing co-occurrence statistics in text corpora. Consequently, distributional semantic models capture relationships between words grounded in patterns of use, rather than linguistic theory. The high dimensionality of vector spaces in LLMs makes it possible to represent complex, multifaceted similarities between words – some of which plausibly reflect conceptual structure and/or inferential roles (Piantadosi & Hill 2022).

4.4. Communicative Grounding

The notion of grounding is also used in relation to another aspect of language, which has to do with communication (Clark & Brennan 1991, Traum 1994). In this context, the 'grounding problem' refers to the problem of establishing *common ground* between speakers in conversation. From this perspective, linguistic communication can be seen as a form of coordinated action that involves collaborating to reach a common understanding of what is said. The process of communicative grounding may involve explicit clarification strategies, such as prompting an interlocutor to repeat or reformulate a garbled or ambiguous statement. It also involves tracking communicative intentions during the exchange, to infer what is communicated from what is said. Such inferences can be made partly on the basis of pragmatic presuppositions, namely, what speakers take for granted or assume to be true when they use certain sentences in the context of the conversation (Stalnaker 2002).

The communicative notion of grounding has also been discussed by natural language processing (NLP) researchers (Brennan 1998, Di Maro 2021, Chandu et al. 2021). For example, Chandu et al. (2021) distinguish between *static* grounding, in which common ground in basic communication between a human user and a virtual agent is automatically established by querying a database (e.g., asking for a weather report); and *dynamic grounding*, in which common ground is more gradually established in back-and-forth communication between a human and a dialogue agent. Bender & Koller (2020) define meaning to be a relation between linguistic expressions and communicative intent, and characterise this relation as a form of grounding.

While communicative and referential grounding are distinct notions, they are related in important ways. Referential grounding anchors language to the world, but the same linguistic expressions may be referentially grounded in different ways for different speakers, due to differences in perception, knowledge, and conceptualisation. This means that calibrating reference in conversation requires communicative grounding – language users must work together to negotiate a common way of connecting language to the world. Without such collaboration, there is no guarantee that speakers really mean the same thing, even if they are using the same words.

4.5. Epistemic Grounding

Finally, the notion of grounding is occasionally used by NLP researchers to refer more minimally to the relationship between linguistic expressions and information stored in knowledge bases (Poon 2013, Tsai & Roth 2016, Chandu et al. 2021, Thoppilan et al. 2022). For example, an expert system may contain a knowledge base of facts and rules belonging to a particular domain. When the system receives user input, it can query this knowledge base to determine an appropriate response. A classic example is MYCIN, an expert system for diagnosing infectious blood diseases (Shortliffe 1977). When given patient symptoms, MYCIN would search its knowledge base of medical rules and facts to suggest possible diagnoses and recommend treatments. In this system, language is grounded in the structured knowledge contained in the expert system's knowledge base. We call this notion *epistemic grounding*.

There are two ways in which current LLMs may involve epistemic grounding. Firstly, they can be hooked up to external databases used to retrieve information (Borgeaud et al. 2022, Thoppilan et al. 2022). Secondly, it has been suggested that even LLMs that are not augmented with retrieval capabilities might be functionally similar to knowledge bases themselves (Petroni et al. 2019). Indeed, they may encode a significant amount of world knowledge in their weights after pre-training, as illustrated by the remarkable performance of state-of-the-art LLMs on a wide range of factual question-answering benchmarks (OpenAI 2023, Bubeck et al. 2023).

4.6. What Grounding for LLMs?

We have identified five distinct notions of grounding that play prominent roles in discussions about meaning grounding in biological and artificial systems. Which of these notions is necessary for LLMs to possess representations and produce outputs endowed with intrinsic meaning? (Recall that intrinsic meaning is meaning that is independent from external interpretation.) To identify the most suitable notion of grounding for this task, we will employ an elimination strategy, examining each notion for its adequacy, beginning with Harnad's1990 proposal of appealing to sensorimotor grounding as a solution to the Symbol Grounding Problem.

Sensorimotor grounding connects conceptual or linguistic representations to sensorimotor representations within the same cognitive system. This sort of grounding can be useful to shed light on some features of the former representations, how they work, and what roles they can play in the economy of a cognitive system. However, sensorimotor grounding falls short in addressing the merry-go-round problem, as it merely grounds one type of representation in another: it does not tackle the question of how sensorimotor representations establish a connection with the world. One might argue that sensorimotor representations are grounded in the world by resembling, to some extent, the entities they represent. However, the success of such an approach relies on an additional factor beyond sensorimotor grounding itself: it is the relevant resemblance relations that are supposed to ground those sensorimotor representations in the world. As sensorimotor grounding does not provide a connection with the world but only with other representations, it does not address the Symbol Grounding Problem. Similarly, this form of grounding does not resolve the Vector Grounding Problem, as grounding vectors in other vectors still leaves open the question of how any of those vectors come to be connected to the world.

The same line of reasoning applies to relational, communicative, and epistemic grounding. Relational grounding anchors linguistic representations in their relationships to other linguistic representations, suggesting that a representation's meaning is determined by its connections to the meanings of other representations within a network of inferential, conceptual, or statistical relations. Although this kind of grounding clearly does not escape the representational carousel, inferential-role semantics may suffice for connecting at least some words to what they are about, e.g., logical connectives, as their meanings are arguably exhausted by the inferences they are involved in (Block 1986, Fodor 1990). Logical connectives are however a rather special case, as they are not supposed to represent entities in the world, but rather abstract relations between any possible entities.[14] Inferential-role semantics can thereby account, at best, for only a very small and special subset of the representations and outputs that biological and artificial systems possess and produce.

Communicative grounding seeks to anchor representations in the shared content of linguistic interactions between agents. However, this shared content relies on the representations within the agents' cognitive systems and the meanings of any public representations used in the exchange. As

[14]Therefore, if an LLM can learn the structure of relations that constitute the meaning of logical connectives and linguistic conjunctions from its training data, then its representations of these terms might be as intrinsically meaningful as they can be.

with sensorimotor grounding, one type of representation is grounded in another, perpetuating the representational carousel without establishing a connection to the world beyond the representations.

Epistemic grounding, on the other hand, aims to ground linguistic representations in knowledge bases. Yet, this approach encounters a familiar obstacle: the representations within the knowledge bases themselves call for an explanation of how they acquire intrinsic meaning. Consequently, epistemic grounding alone fails to provide a comprehensive solution to the Grounding Problem, as it does not address the question of how representations come to possess intrinsic meaning to start with.

In sum, the form of grounding that underlies the other four types and is most relevant to addressing the Grounding Problems is what we have called referential grounding. Referential grounding anchors a representation – be it a symbol, a map, or a vector – in the world itself, rather than in other representations. It stands as the sole form of grounding that escapes the representational merry-go-round and connects a representation to whatever worldly entity or property it is about.[15] Consequently, to determine whether Generative Pre-Trained Models such as LLMs manipulate and produce representations and outputs with intrinsic meaning, we must assess whether they can achieve referential grounding. To do so, we first need to examine the requirements for a representation to be referentially grounded.

5. Theories of Representational Content: Causes and History

Philosophers of cognitive science have dedicated considerable effort over the past several decades to developing theories that explain how mental and cognitive states, such as beliefs, desires, and neural activations, can be *about*, or *mean*, entities and properties in the world. What mental and cognitive processes are about, or mean, is their 'representational content', or 'content' for short. The philosophical project of developing theories of representational content has the aim of illuminating how meaning can emerge in a world that otherwise seems devoid of it. After all, natural objects such as rocks and lakes generally do not possess internal states endowed with meaning and, by themselves at least, do not mean anything.

Some internal representations are basic, in the sense that their meaning, or content, does not derive from other states already endowed with meaning. They can be seen as the building blocks of more complex representations, which draw some of their meaning from those basic representations. For instance, 'edge detectors' in primary visual cortex (V1) are good candidates for being basic internal representations, whereas 'shape detectors' in downstream visual areas are not, since their contents derive partly from the contents of the basic edge detectors. Similarly, words and sentences are not basic in this sense, since their meaning derives from other, more basic representational states, such as beliefs, desires and intentions – themselves complex representations.

The distinction between basic and complex representations cuts across the distinction between intrinsic and extrinsic meaning. As a reminder, states with intrinsic meaning are those that carry meaning independently of an interpreter with mental states, such as a human, attributing meaning to them. Both the representations in primary visual cortex and the words and sentences humans utter are cases of representation with intrinsic meaning, though the former are basic while the latter are complex. By contrast, words randomly formed by the movement of ants in the sand, outputs from a random sentence generator – and, according to Bender & Koller (2020), the outputs of LLMs – only have extrinsic meaning: they are merely shapes that require human interpretation to convey any meaning at all.

[15] Or almost: see the caveat about inferential-role semantics and logical connectives above.

We cannot here do justice to the rich and nuanced interdisciplinary research that has been carried out for the past several decades on the notion of cognitive representation.[16] What matters for our limited purposes is that, despite some disagreements on the finer details, philosophers of cognitive science have largely reached a consensus on the key factors that enable internal states in a system to become representations, both basic and complex (Shea 2018, Coelho Mollo 2022).

The first, most fundamental point of agreement is that an account of basic representational content must identify one or more natural relations between internal states and external entities and properties responsible for making the former be about the latter. These natural relations must not be ones that already involve states with meaning. Since the goal is to account for the emergence of meaning in the world, theories of basic content must exclusively rely on meaning-*less* states and processes to avoid vicious circularity. Much of the work in this area has been dedicated to identifying those relations that can yield fruitful content ascriptions and contribute to explaining the behaviour of cognitive systems.

Two natural relations are widely considered to be central to endowing basic internal states with meaning: *causal-informational* relations, and *historical* relations. In general terms, causal-informational relations occur when one state of the world conveys information about another due to a causally-generated correlation between the two states. A classic example is the relationship between smoke and fire: smoke is correlated with fire because it is caused by it, so the presence of smoke carries information about the presence of fire. Neurons in brain area V1 convey information about edges in the perceived visual scene, as their activation correlates (through a relatively complex causal chain) with the presence of edges in the world in front of the perceiving subject.

Causal-informational relations capture the main rationale for why systems possess basic internal representations: they allow systems to be sensitive to specific features of their environments – those for which internal states stand in causal-informational relations to. However, the mere existence of a causal-informational relation is insufficient to provide an account of basic representational content. An additional requirement often mentioned is that the causal-informational relation be exploited, that is, used by the system due to its information-carrying nature (Shea 2018). In other words, a system must use the presence of smoke to obtain information about the presence of fire, form a belief about there being a fire, guide fire-related behaviour, and so on, for smoke to be a viable candidate for representing fire for that system.

Similarly, the information about the world that edge detectors in V1 carry into the mammalian brain must be used downstream for further processing – such as detecting food sources – so as to guide appropriate behaviour. To be a representation necessarily involves playing a certain role: that of conveying information about the world and being used by the system in virtue of that. This further requirement complements the explanation for why internal representations are useful to systems: they are used by the system as sources of information about the world. It also helps to avoid the risk of trivialising the notion. Causal-informational relations are abundant in nature: every effect carries information about its cause. If every such relation should suffice for being a representation, representations would be so extremely trivial that their special explanatory power for cognition would be lost.

Exploited causal-informational relations alone, however, are still not enough to base an adequate theory of representational content, for another crucial feature of internal representations is missing: their capacity for truth or falsehood, correctness or incorrectness. To be a representation inherently involves the possibility of *misrepresenting* – the possibility of getting things wrong and failing to fit to the world accurately. V1 neurons may become active when no edge is present, as in the famous

[16] See Neander (2017), Millikan (2017), Shea (2018) for recent book-length defences of the main contenders today.

Kanizsa's triangle illusion, thus conveying information to the system that, if not corrected by other subsystems, can lead to inappropriate behaviour (such as trying to touch the non-existent edge). Representations involve a kind of descriptive normativity (Neander 2017). That is to say, representations can represent successfully or unsuccessfully, better or worse. This descriptive normativity extends to external representations, such as sentences, and meaning more generally. They are kinds of things that can be true or false, accurate or inaccurate, appropriate or inappropriate. The rationale for systems to possess representations involves that they be accurate enough, enough of the time. Otherwise, representations would lead systems astray and hinder appropriate behaviour, rather than the other way around. Generally, representations have the *function* to represent correctly and accurately (to a degree of required precision).[17]

As the notion of representation is central for cognitive (and computer) science, the notion of function is central to biology. This has led philosophers of biology to develop scientifically-grounded, objective theories of function Wright (1973), Millikan (1989), Neander (1991), Godfrey-Smith (1994), Garson (2019). If functions exist out there in the world, they must be a matter of natural features and relations, rather than human interpretation. The dominant view of functions is the selected-effects theory Millikan (1989), Neander (1991), Garson (2019). On this view, functions are ensembles of the historical causes that explain the persistence of a certain trait or feature in a system. For instance, hearts today have the function to pump blood because past hearts, by pumping blood, causally contributed to the survival and reproduction of the systems having them. Pumping blood is what explains, causal-historically, why hearts still exist – and it is thereby their function. The relevant causes can be at the level of the species, typically through natural selection; or at the level of the individual, through selection-like learning processes during development.

The most influential views on representation incorporate the selected-effects theory to account for internal states having the function to carry information about the world and to guide appropriate behaviour Neander (2017), Millikan (2017), Shea (2018), Coelho Mollo (2022). The guiding idea is that, through selection and/or learning, states and processes in cognitive systems come to acquire functions they are supposed to perform. For example, neurons in V1 have the function of detecting edges in the visual scene because they were selected for this purpose during evolution via natural selection; and such selection occurred due to the causal-informational relations they maintain with actual edges in the world. When these neurons successfully detect edges, they fulfil their function, representing correctly; when they fail to fire in the presence of an edge or fire in the absence of one, they malfunction, misrepresenting. In brief, internal representations are states that carry information about the world, have the function to do so, and are appropriately used by the system in virtue of the information they carry.

Existing theories of basic representation vary in the emphasis they place on causal-informational or historical relations, as well as in the specific types of such relations they prioritise. Additionally, they may diverge on the inclusion of further relations, such as structural resemblance relations Shea (2018). Nevertheless, all of these theories, in one way or another, assert that a combination of these relations is central to the emergence of meaning in the world. To possess intrinsic meaning, at a minimum, is to stand in a specific set of causal-informational and historical relations with the thing in

[17] In some special cases, systematic misrepresentation can be a sensible price to pay. If the stakes are high enough, it makes sense for representations to err on the side caution: if you are trying to detect predators, it makes sense to produce many false-positive (misrepresenting predators as present when there are none), rather than even one, potentially fatal false-negative.

the world being represented. Theories of content, therefore, provide the essential ingredients that determine referential grounding, enabling internal states to hook onto things in the world.[18]

6. Grounding Generative Pre-Trained Models

Philosophical theories of representation identify sets of relations that imbue representations with content, giving them intrinsic meaning. These primarily include causal-informational and historical relations. Our aim is to show that Generative Pre-Trained Models, such as LLMs, can have internal states and generate outputs with intrinsic meaning, i.e., that are referentially grounded in the world.[19] To begin, we will examine whether appropriate causal-informational relations exist between LLM representations/outputs and entities and properties in the world.

6.1. Causal-informational relations

Although LLMs are trained exclusively on linguistic data, thus precluding any direct causal-informational relationship with extra-linguistic entities, the linguistic data itself is generated by humans who stand in causal-informational relations to worldly entities and properties. Those data are shaped by humans' interactions with the world. Among its various functions, language enables humans to express facts about the world, as well as their experiences, emotions, and ideas concerning it. The extensive corpus of text upon which LLMs are trained bears the 'imprint' that the world has left on language and linguistic expression.

A strict separation of form and meaning has been a central assumption in much of linguistics, from Ferdinand de Saussure's principle of the "arbitrariness of the linguistic sign" (de Saussure 1916) to contemporary formal theories of syntax and semantics (Chomsky 1957, Montague 1970). However, this notion has faced growing criticism since the 1980s, especially in light of empirical investigation of the iconicity of language (Perniss & Vigliocco 2014). In this context, iconicity refers to a non-arbitrary relationship between the form of a linguistic sign and its referent, based on some similarity or analogy (Haiman 1985). The nature of this similarity relation can vary: imagistic iconicity involves a perceptual resemblance between a word and its referent (e.g., in onomatopoeias), while diagrammatic iconicity involves a structural resemblance .

Diagrammatic iconicity is pervasive in linguistic structures. For example, the ordering of events in narrative sequences tends to reflect closeness and succession in time (e.g. "Veni, vidi, vici"); what is nearest to the speaker in a literal or metaphorical sense tends to be mentioned first (e.g., "here and there"); elements whose referents are closely related in some way or other tend to be adjacent in sentences (principle of adjacency); the use of subordination reflects causality or conditionality between states-of-affairs (e.g., "If you study hard, you will pass the exam"), and so on (Van Langendonck 2010). The pervasiveness of diagrammatic iconicity in linguistic structures suggests that the way in which linguistic signs are organised and structured is not arbitrary, and reflect some ways in which the world is organised and structured.

While diagrammatic iconicity is not typically characterised in distributional terms, patterns of co-occurrence in language reflect stable statistical regularities in the way in which linguistic elements are used together. Certain words tend to occur together more frequently than others, and certain grammatical constructions tend to co-occur with particular semantic categories. These patterns of

[18] Philosophical discussions about cognitive representation have traditionally given considerable weight to metaphysical worries regarding the determinacy of content. Once seen instead as scientific, explanatory theories, we take such worries to be misguided (cf., Shea 2018, Coelho Mollo 2022).

[19] For recent discussions that apply theories of representational content to LLMs, see Harding (2023), Grzankowski (2024), Goldstein & Levinstein (2024), Chalmers (2025).

co-occurrence reflect the underlying structure of the world that speakers are trying to describe. Thus, *contra* Bender & Koller (2020), form and meaning are far from being so easily separable, especially at the level of general patterns of language use.

The patterns of language use derived from the training corpus reveal important aspects of the world. They embed information about patterns of interaction between humans and the world, human ways of categorising their environments, and the selection of features that are more or less salient to them. Indeed, a plausible hypothesis for LLMs' success in a wide range of tasks is that they have induced general semantic patterns from their training data, which closely resemble what could be termed semantic 'laws'.[20] These laws are not merely formal or syntactic; rather, they tap into the intricate relationships between linguistic form and the world it describes.[21]

In short, LLMs are trained on data that is shaped by humans and their interactions with the world. LLMs stand in no direct causal connection to worldly entities, but they do stand in indirect ones, mediated by the humans who generated the training data. Importantly, while the complete lack of direct causal relations to the world sets LLMs apart from humans, reliance on indirect causal relations does not. Many human representations are also acquired by indirect, mediated ways. Indeed, testimony and social learning underlie a considerable amount of the internal representations humans come to learn, and are considered to be key processes that help explain cultural knowledge and cultural evolution – in their turn central factors for the species' success Tomasello (2009), Sterelny (2014), Heyes (2018).

These include representations of entities one has not had any direct causal contact with. For example, most people have never been in direct causal contact with enriched uranium, but that does not entail that their thoughts and statements about uranium lack referential grounding – they are indeed about uranium. Representations of culturally constructed notions, such as 'democracy' or 'university', are also acquired this way – they rely on complex sets of social practices and conventions that are transmitted not by direct causal contact with documents or buildings, but rather by social learning.

In a way akin to many human representations, the internal representations in LLMs are indirectly connected to the world through testimony and 'social' learning. Since their training data contains innumerable instances of human communication, LLMs should be seen as an additional link in the chain of social information transmission that endows representations with causal-informational relations to worldly entities, and thus partly underpins intrinsic meaning.

In a sense, thereby, the causal-informational relationships between LLMs and the world are dependent on humans. This dependence is however distinct from the one that drives the Vector Grounding Problem. The Problem is not *how* the representations of connectionist models come to acquire intrinsic meaning, but *whether* these representations have intrinsic meaning at all – instead of being merely ascribed meaning by external interpreters. To take once again the example of enriched uranium, LLMs find themselves in a position similar to most humans: their representations and outputs about enriched uranium intrinsically refer to enriched uranium – irrespective of any interpreter's gloss – even though the process enabling such reference depends on a chain of transmission that involves other systems with intrinsic meaning. In our terminology, such internal states are non-basic representations with intrinsic meaning.

In sum, LLMs possess the necessary elements to fulfil the causal-informational requirement for referential grounding. They stand in complex, mediated causal-informational relations to the world.

[20] For some preliminary work on the potential form of such semantic laws, see Wolfram (2023).

[21] See Merrill et al. (2024) for concrete evidence that LLMs can induce entailment relationships between sentences from co-occurrence statistics.

These connections are mediated since they pass through human linguistic production, where language primarily functions as a tool to express aspects of the world.[22] They are complex due to the intricate causal-informational relationships between the patterns of language use, the semantic 'laws' that govern them, the structure of the world they have (culturally) evolved to capture, and the causal chains of communication and testimony that produce the learning data – of which the LLM itself can be considered a link. In light of these considerations, we contend that Generative Pre-Trained Models, such as LLMs, can meet the causal-informational criteria necessary for possessing referentially grounded representations and outputs.

6.2. Historical Relations

Although it is important to differentiate between causal-informational and historical relations when considering referential grounding, these aspects are often intertwined in real-world cases. The causal-informational relations previously discussed with respect to LLMs possess a historical dimension, as they depend on patterns of interaction between humans and their environments. Likewise, in many of the interesting cases pertaining to human cognition, the meaning-endowing causal-informational relationships can be traced back to past episodes of learning. That said, as discussed in section 5, finding causal-informational relations between LLMs and the world is the easiest part of our investigation, given how abundant such relations are. Much more challenging is to show that LLMs have internal states that not only carry information about worldly entities, but that also have the *function* to carry such information – so that the LLM can use it in generating appropriate outputs.

Selection processes and learning are the historical mechanisms that theories of representation commonly invoke to explain how representations come to have the relevant function. These processes explain how a system acquires a specific function or purpose, in this case, the production and/or use of representations. Such functions clarify how cognitive systems can represent *accurately* – when they produce and use representations that fulfil the relevant function – as well as *misrepresent*, when the representations produced and/or used are inappropriate for the task. While it is widely accepted that Generative Pre-Trained Models undergo a learning-like process during training, it is less clear whether they undergo the relevant kind of selection process. It can be argued that they do, at least to the same degree as other human artefacts (Artiga 2016, Coelho Mollo 2019), given that numerous design and implementation choices form an integral part of creating artefacts such as AI systems.

The crux of the matter, however, lies in determining whether LLMs stand in historical relations to the *right* sorts of things, namely the entities and properties in the world that could ground their representations Butlin (2021). At first glance, the situation appears analogous to that of causal-informational relations: LLMs stand in historical relations to the world, albeit in a mediated manner, through the human-generated data used for training them. The critical question is: can the internal states and outputs of LLMs acquire the function to represent extra-linguistic entities, despite their being trained only on language? In other words, the question is whether LLMs can acquire the sort of world-involving functions needed for referential grounding, and which underlie the descriptive normativity that is an essential feature of meaning.

To address this question, it is important to distinguish between the proximate and ultimate functions performed by LLMs during training and inference. The proximate function of LLMs is plainly

[22] This analysis would extend, *mutatis mutandis*, to multimodal inputs, which are also mediated by human interactions with the world (e.g., photographs are captured by humans with a camera). However, one might argue that some worldly properties instantiated by visual samples in a multimodal dataset, such as shapes, can be accessed more directly than they could through the mediation of language. This is because images (and samples in other sensory domains) are entirely iconic, while language only exhibits limited patterns of iconicity. We discuss whether multimodal models fulfil the historical requirement for representational content in subsection 7.3 below.

next-token prediction: predicting the next token in a text sequence, given the context provided by that sequence. This proximate function is infra-linguistic, and so are the standards for its fulfilment or failure, namely, whether the LLM predicts the correct tokens reliably enough. While such function may be suitable for underpinning *relational grounding*, it appears inadequate for contributing to the world-involving normativity necessary for *referential grounding*. In contrast, ultimate functions refer to the broader, more abstract goals that the model is tasked with achieving through next-token prediction. These functions can be world-involving in principle, such as generating text that accurately describes real-world states of affairs, or solving problems based on real-world information. In what follows, we will consider several potential sources for world-involving ultimate functions in the training and use of LLMs.

Our general argument goes as follows. The training process of LLMs is a form of selection – through learning – that stabilises the functional roles of the model's internal states. This process selects for internal states whose ultimate functional role extends beyond mere linguistic prediction to include extra-linguistic, world-involving functions. This is most evident in models that undergo fine-tuning to satisfy human preferences, since such fine-tuning explicitly involves selecting internal states that increase the probability of outputs satisfying world-involving norms, such as factual accuracy. However, we will also present evidence suggesting that pre-training alone can, in some contexts, select for internal states with world-involving content, albeit in a more indirect way.

If LLMs indeed have internal states with extra-linguistic intrinsic meaning, and if those states causally influence the generation of linguistic outputs, then those outputs are thereby referentially grounded in the world. This would mean that at least some outputs of LLMs are intrinsically meaningful in the sense discussed earlier: their meaning does not depend on the interpretations projected onto them by human users.

The key premise in this argument – that the training process of LLMs can select for internal states with extra-linguistic functional roles – requires detailed defence. In what follows, we offer two complementary arguments supporting this premise. The first, which we call the *argument from post-training*, focuses on how preference tuning methods explicitly establish world-involving functional roles for LLM internal states. The second, which we call the *argument from pre-training*, examines how even the basic next-token prediction learning objective might, under certain conditions, implicitly select for ultimate world-involving functions.

6.2.1. The Argument from Post-Training

The training process of modern LLMs, particularly those deployed in consumer-facing chatbots like ChatGPT or Claude, involves two distinct phases: *pre-training* and *post-training*. Pre-training involves training the base model from scratch on a next-token prediction objective across vast corpora of text. By contrast, post-training adapts this base model to behave in ways that go beyond generating merely plausible text completions, making LLMs more useful for various downstream tasks and applications. The post-training phase typically includes two types of fine-tuning: *instruction tuning*, and *preference tuning*.

Instruction tuning is a form of supervised fine-tuning on curated instruction-response pairs (Ouyang et al. 2022). Given an instruction, the model is trained to generate the desired output by minimising the difference between its predictions and the target responses in the instruction dataset. This teaches the model to generate appropriate responses that follow user instructions rather than simply generating plausible completions of text sequences.

Preference tuning, by contrast, trains LLMs to generate outputs that better align with selected normative criteria, such as helpfulness, harmlessness, and honesty or factual accuracy (Askell et al.

2021). Unlike standard supervised fine-tuning, which only demonstrates what to do, preference tuning explicitly trains models to distinguish between better and worse responses according to these criteria. This is typically implemented by training on pairs of responses where human evaluators have indicated a preference for one over the other.

Reinforcement Learning from Human Feedback (RLHF) is the most well-known preference tuning method (Christiano et al. 2017). RLHF begins by collecting human preferences over model outputs for a diverse set of prompts. These preferences are used to train a reward model that predicts how humans would evaluate various responses. The LLM is then fine-tuned using reinforcement learning to maximize this reward function while remaining reasonably close to its original behaviour. Another increasingly popular approach is Direct Preference Optimization (DPO), which achieves similar results without explicitly modelling the reward function (Rafailov et al. 2024). Instead, DPO directly optimises the model to assign higher probabilities to preferred responses and lower probabilities to rejected ones.

When an LLM is optimised according to human preferences that incorporate extra-linguistic normative standards like factual accuracy, this creates a selection process that reinforces specific activation patterns based on their relationship to worldly states. These internal states acquire a specific functional role: to track relevant aspects of the world that enable the model to generate outputs that satisfy human preferences for truthfulness.

To illustrate how preference tuning establishes world-involving functions, consider a toy example focused solely on factual accuracy using DPO. We begin with a language model that has undergone both pre-training and instruction tuning. This model can follow instructions reasonably well, but has not been specifically optimised for factual accuracy. To align the model with the norm of factual accuracy, we collect a dataset of human preferences where judges rank responses based solely on which one contains more truthful information and fewer falsehoods.

When we apply DPO to this model, it adjusts the model's parameters to increase the probability of generating factually accurate responses and decrease the probability of generating less accurate ones.[23] Through this process, the model's internal states are selected specifically because they enable the model to satisfy the norm of factual accuracy, thereby providing a standard of success that depends not on linguistic patterns alone, but on the correspondence between the generated text and states of affairs in the world.

This functional role provides another piece of the story – in addition to causal-informational relations – of how LLM internal states can be referentially grounded, and thus overcome the Vector Grounding Problem. The internal states of the preference-tuned LLM represent aspects of the world because they were selected specifically for their ability to carry information about those aspects in service of generating truthful outputs. When the model succeeds in generating a factually accurate response, its internal representations are fulfilling their function of tracking relevant worldly states.

Importantly, the accuracy of the model is not what determines whether its internal states are referentially grounded. Even a partially inaccurate model prone to occasional hallucinations can have internal states with world-involving content, provided those states were selected for their ability to carry information about aspects of the world. What matters for grounding is not perfect accuracy, but rather the presence of a representational function with world-involving correctness conditions. Preference tuning thus provides a clear mechanism by which LLMs can acquire internal states with

[23] Note that DPO also incorporates a constraint that prevents the model from deviating too far from its original behaviour, to maintain the right balance between preference satisfaction and fluency.

extra-linguistic functional roles, insofar as RLHF or DPO incorporate factual accuracy as an explicit criterion over rewarded preferences.[24]

While we focus on factual accuracy, all three common norms in preference tuning – the "3 Hs" of helpfulness, harmlessness, and honesty – are arguably world-involving. Helpfulness requires generating outputs that genuinely assist users with their goals, which depends on real-world task success conditions rather than mere linguistic coherence. Harmlessness involves avoiding outputs that could lead to real-world negative consequences, requiring sensitivity to causal relationships between linguistic outputs and potential worldly outcomes. Even instruction tuning alone could potentially establish world-involving functions, as successfully following instructions often requires tracking aspects of the world that would make the instruction completion successful. Many instructions implicitly contain success conditions that depend on extra-linguistic factors. Be that as it may, factual accuracy provides the most direct and straightforward case of world-involving content, where the connection between internal states and worldly conditions is the most evident. It offers the clearest case for how LLM internal states can be referentially grounded, and thus endowed with intrinsic meaning.

6.2.2. The Argument from Pre-Training

We now turn to a more controversial claim: that pre-training alone can, in certain contexts, establish world-involving functions.

During pre-training, LLMs are explicitly trained to minimise the negative log-likelihood of the next token given previous tokens – an objective limited to the statistical distribution of tokens in the training corpus.[25] At first glance, this appears to be a purely infra-linguistic task with no connection to the world. However, as we have seen in subsection 6.1, the patterns of language use in the training corpus are not arbitrary; they reflect systematic regularities in how humans use language to communicate about the world.

From an information-theoretic perspective, there are vastly more ways to linguistically express world states than there are world states themselves. For next-token prediction, two strategies are possible: direct memorisation of observed sequences, or latent variable factorisation that learns to represent underlying causal factors. The parameter efficiency of the latter strategy relative to the former scales with the ratio of linguistic expressions to world states – typically a very large number. This creates a strong selection pressure for internal states that track latent variables, as they enable more efficient prediction across diverse linguistic contexts.

The Information Bottleneck principle provides a formal framework for understanding this process (Tishby & Zaslavsky 2015). Neural networks learn to extract approximate, minimally sufficient statistics of the input with respect to the output. This involves finding a maximally compressed mapping of the input that preserves as much information as possible about the output. In language models, this connection is even more fundamental. As shown by Delétang et al. (2024), minimising the next-token prediction loss is mathematically equivalent to finding the most efficient way to compress tokenised information. For complex data distributions generated by structured world domains, the most efficient compression involves factorising the problem according to the causal structure that generated the data.

[24] At least in those domains where such factual accuracy has been emphasized during preference tuning.

[25] See McCoy et al. (2024) for an approach to studying LLMs, which they call the "the teleological approach", that takes as its starting point this proximate function. They show that LLM behaviour is strongly shaped by the distributional statistics of the training dataset. As we argue here, however, this is compatible with LLMs also, at least in some cases, fulfilling ultimate, world-involving functions. Sensitivity to distributional patterns can be a means to fulfilling ultimate goals, given all the worldly information embedded in such patterns.

Research in mechanistic interpretability provides concrete evidence that models pre-trained on next-token prediction can develop internal representations with extra-linguistic content at least in some domains. One particularly interesting case is Othello-GPT, a transformer model trained solely to predict moves in Othello game transcripts (Li et al. 2023). Despite receiving no explicit information about Othello's rules or board structure, this model was found to develop internal representations that encode the current game state in a player-centric manner. Nanda et al. (2023) showed through intervention studies that specific directions in the model's activation space encode board positions as "Mine," "Yours," or "Empty." When they intervened on these representations by 'flipping' a tile from 'yours' to 'mine' in the model's activation space, the model's predictions changed in ways consistent with the altered board state. This suggests not only that the model encodes the board state, but that this encoding plays a causal role in determining its predictions of legal next moves.[26]

The Othello-GPT example illustrates how selection processes during pre-training can establish ultimate functional roles with extra-linguistic correctness conditions. During training, the model developed specific directions in its activation space that linearly encode the board state from the perspective of the current player. These linear representations emerged because they enabled the model to accurately predict legal moves across diverse game contexts. Their function became to carry information about the current state of the Othello board, with accuracy conditions dependent on the actual game state rather than merely on linguistic patterns. This example is particularly compelling because game transcripts are systematically constrained by the rules of Othello – to predict legal moves effectively, the model developed a linearly decodable internal model that tracks the underlying game state.

The case of LLMs trained on diverse corpora is undoubtedly more complex than Othello-GPT. Games represent a special class of domains with respect to the grounding problem. Unlike open-ended conversation, where success criteria may be fuzzy and context-dependent, games typically have explicit, formalised rules that define success and failure categorically. These rules create a clearly bounded domain with a fixed ontology (game pieces, positions, legal moves), and deterministic success conditions. By contrast, general LLMs encounter messy, open-ended data with few rigid underlying rules. To the extent that language models may acquire internal states with world-involving content, similarly to Othello-GPT with respect to the game world, this presumably only happens in limited domains and contexts, in which next-token prediction on text sequences provides the right kind of learning signal. That said, there is converging evidence from mechanistic interpretability research that LLMs have internal representations that carry information about discourse entities and their relations, and that such representations are causally relevant to the generation of outputs (e.g., Li et al. 2021; see Millière & Buckner forthcoming for a review).

A useful perspective for understanding how pre-training can establish world-involving functional roles is to view it through the lens of meta-learning. Meta-learning – often characterized as 'learning to learn' – is the process by which a model improves its learning capabilities through experience across multiple tasks, essentially learning how to learn more efficiently. The pre-training of LLMs can be viewed as a form of meta-learning because it trains the model on a diverse distribution of tasks implicitly contained within the next-token prediction objective. When predicting the next token in various contexts across a large corpus, the model encounters many implicit 'tasks', ranging from syntactic completion to factual recall, reasoning, and various domain-specific problems. This defines a bi-level optimisation structure characteristic of meta-learning: an outer loop (the training process itself) that optimises model parameters, and an inner loop (the forward pass during inference) that applies these learned parameters to solve new tasks on the fly (Wu & Varshney 2024). This helps explain why pre-trained LLMs can rapidly adapt to new tasks from just a few examples at inference time,

[26] For a longer philosophical discussion of causal interventions used to understand interrnal states of Othello-GPT, see Millière & Buckner (forthcoming).

a capacity known as *in-context learning* (ICL): they have essentially learned optimisation strategies that generalise across the diverse task distribution encountered during pre-training, enabling efficient transfer to novel contexts during inference.

Many different approaches to understanding in-context learning have been proposed. One prominent theory suggests that ICL involves implicitly optimising over latent task objectives. For example, von Oswald et al. (2024) showed that autoregressive transformers trained on sequence prediction tasks develop gradient-based optimisation algorithms in their forward pass. This suggests that when pre-trained LLMs encounter a series of examples in their context window, they can perform an implicit optimisation process over the latent task space. In other words, they can identify the underlying task structure from the examples, and construct an internal objective that captures the task function. When the in-context examples reflect an extra-linguistic task, the implicit objective function defined by the task may also be extra-linguistic.[27] Optimising this implicit function during the forward pass should dynamically select activation patterns that increase the likelihood of outputs satisfying the world-involving task constraints.

In summary, the argument from pre-training suggests that even without post-training methods like preference tuning, language models may develop internal representations with ultimate world-involving functions through the selective process of next-token prediction over structured text data, at least in some limited domains. This can occur because inducing the latent causal structure of the data provides an informationally efficient solution to the prediction task across diverse contexts. Evidence from mechanistic interpretability studies supports this view at least in toy domains, showing that Generative Pre-Trained Models can develop internal states that carry information about extra-linguistic properties and causally influence outputs. Furthermore, the meta-learning perspective on in-context learning suggests that pre-trained LLMs can implicitly optimise ultimate extra-linguistic task functions at inference time without parameter update.

7. Implications

7.1. Parameter-identical duplicates

The historical component of our account of referential grounding invites consideration of a classic philosophical objection. The 'swampman' thought experiment envisions a perfect duplicate of a human spontaneously formed by random chance (in the original thought experiment, by a lightning strike on a swamp), and thus without any causal history. We can construct an analogous thought experiment for LLMs:

> **Swamp LLM.** Suppose a language model M_1 is trained normally, undergoing both pre-training on next-token prediction and post-training with instruction tuning and preference tuning. Now imagine that an untrained language model M_2 is randomly initialised and, by extraordinary coincidence, happens to have exactly the same parameter values as the fully trained M_1, despite never having been exposed to any training data.

If M_1 achieves referential grounding on our view, does M_2 also achieve it? Our answer aligns with the standard response to the swampman scenario: despite being parameter-identical to M_1, M_2 would lack referential grounding. First, the internal states of the swamp system, at least at the moment of creation, would lack any causal-informational relations to the world, given that, by hypothesis, the

[27] See, for example, Patel & Pavlick (2022)'s experiment which involves prompting the model in a few-short learning regime to learn mappings from colour terms like 'cyan' to points in RGB colour space such as `[0, 100, 100]`. This is an instance of a world-involving function, having a standard of success and failure directly contingent on how the world is.

system has no causal history. Moreover, the absence of a selection and learning history means that the internal states of the swamp system lack any functions, including functions to carry information about the world Shea (2018). Consequently, there is no principled way to determine which outcomes count as successes versus failures for a swamp system. M_2 might produce outputs that appear meaningful to human interpreters, but without the selective history that shaped M_1, there is no objective basis for treating some of its outputs as 'correct' and others as 'incorrect'. The parameter values themselves, divorced from the process that selected them, cannot ground representational content.[28]

It is worth noting, however, that referential grounding could be established relatively quickly if M_2 were subsequently fine-tuned. Even a brief learning history would begin to establish the relevant world-involving functions Shea (2018). What matters for grounding is not the parameter values per se, but the historical process through which those parameters were selected to perform specific functions.

We can also consider a variant of this thought experiment:[29]

> **Lucky Pre-Training.** Suppose M_1 is trained normally, undergoing both pre-training and post-training. M_2 undergoes only pre-training on next-token prediction, but through some extraordinary coincidence, ends up with exactly the same parameter values as the fully trained M_1.

This scenario differs importantly from the standard swampman case, as M_2 does have a learning history – it just differs from that of M_1. The answer to whether M_2 achieves referential grounding depends on whether one accepts our argument from pre-training in Section 6.2.2.

If one accepts that pre-training alone can, in some cases, establish world-involving functional roles (as we argued), then M_2 could achieve referential grounding despite skipping the post-training phase. However, the fact that M_2 happens to have identical parameter values to M_1 would be a mere coincidence irrelevant to the grounding question. Indeed, its referential grounding would be limited to that achieved via pre-training, while the internal states that drew their grounding from preference tuning in M_1 would be ungrounded in M_2.

These thought experiments highlight the crucial role that historical relations play in referential grounding. The internal states of a system, no matter how complex or functionally organised, cannot have representational content without the appropriate history. At the same time, our position avoids an overly restrictive view of what learning histories can establish referential grounding, allowing that different training regimes might enable a system to overcome the Vector Grounding Problem through different causal-historical paths.

7.2. The impact of fine-tuning

In a recent paper, Grindrod (2024) raises an objection to our argument from post-training. Grindrod acknowledges our claim that fine-tuning with human feedback introduces world-involving functions that could in principle establish the necessary historical relations for referential grounding. However, he objects:

> The problem with this idea is that it would be odd to claim that a relatively marginal procedure that is designed to optimize the performance of a LLM on specific tasks ends

[28] This scenario is somewhat akin to the example, mentioned above, of ants forming the shape of words in the sand by pure chance; or of chimps eventually reproducing a Shakespeare play by randomly hitting keys in a typewriter. Such outputs have no intrinsic meaning, as they were not produced in the appropriate ways, though referentially grounded beings, such as ourselves, may extrinsically impose meaning on them.

[29] We thank an anonymous referee for raising this scenario.

> up being the difference between intentionality and its lack. The reason, simply put, is that the fine-tuning procedures – whether RLHF, supervised training, or something else – are not the reason why we have such impressive LLMs today. (Grindrod 2024, p. 71)

Grindrod's objection rests on the premise that it would be implausible for a "relatively marginal procedure" to be the critical factor determining whether LLMs possess intrinsic meaning. He points out that pre-trained models without preference tuning already demonstrate impressive capabilities, suggesting that whatever makes LLMs impressive cannot be the same factor that provides them with intrinsic meaning.

This objection misunderstands the purpose of our appeal to preference tuning. We are not claiming that preference tuning explains the impressive capabilities of LLMs or their ability to produce contextually appropriate outputs. Rather, we are specifically addressing how LLMs can overcome the Vector Grounding Problem – how their internal states can possess intrinsic meaning by being referentially grounded in the world. These are distinct questions.

Intrinsic meaning does not determine the complexity or impressiveness of outputs, neither in biological nor in artificial systems. Many relatively simple biological systems possess internal states with intrinsic meaning, while many complex systems might generate impressive outputs without intrinsic meaning. The quality of outputs and their status as intrinsically meaningful are orthogonal issues.

What matters for referential grounding is not how much a particular training procedure contributes to performance quality, but whether it establishes the right kind of functional relations between internal states and worldly entities. Even if preference tuning procedures make only marginal contributions to an LLM's overall capabilities, they can fundamentally alter the normative conditions under which the model operates by introducing world-involving correctness standards.

7.3. Multimodality

The preceding discussion has identified two potential sources of referential grounding for LLMs: (1) preference tuning with human feedback that incorporates world-involving criteria such as factual accuracy, and (2) pre-training on next-token prediction, which may select for internal states with world-involving functional roles in limited domains where this is the most efficient way to achieve good prediction accuracy.

When we turn to multimodal and embodied systems, we need to examine whether and how these sources of referential grounding extend to such systems. Our analysis leads to a surprising conclusion: multimodality and embodiment are neither necessary nor sufficient conditions for referential grounding in artificial systems, though they can complement and potentially enhance referential grounding when the right kind of learning process is available.

Let us first consider a simple multimodal model like CLIP (Radford et al. 2021). CLIP is developed by jointly training an image encoder and a text encoder to predict the correct pairings of images and text captions using a contrastive learning objective. During training, CLIP receives batches of image-text pairs and learns to maximize the similarity between embeddings of matched pairs while minimizing the similarity between embeddings of unmatched pairs. This training approach allows CLIP to develop representations that capture semantic relationships between visual and linguistic content. Through contrastive pre-training with image-caption pairs, CLIP achieves what we called *sensorimotor grounding*, albeit to a very limited extent compared to biological organisms – namely connecting linguistic vector representations to pictorial vector representations. CLIP also arguably achieves some degree of *relational grounding* in the purely linguistic domain, although there is

evidence that contrastive learning from image-caption pairs severely limits the degree to which this kind of model can induce syntactic structure compared to text-only LLMs (Yuksekgonul et al. 2023).

As far as causal-informational relations go, the account we gave for LLMs should also apply to models like CLIP: since both images and their captions are shaped by human interactions with the world, CLIP should stand in complex, mediated causal-informational relations to the world. The images in CLIP's training data were produced by humans using cameras to capture real-world scenes, and the captions were written by humans to describe those scenes. This establishes an indirect causal chain connecting CLIP's internal representations to worldly entities and properties. However, the contrastive learning objective optimises purely for matching images with their captions, without establishing clear world-involving functional roles for CLIP's internal states. There is no obvious selection pressure for internal states that track features of the world beyond what is necessary for the matching task, and no clear normative standard for success beyond the matching task itself. Thus, whether a model like CLIP can achieve referential grounding through the pathway outlined in the argument from pre-training is doubtful.

By contrast with CLIP, some multimodal models are built with a pre-trained language model backbone. For example, typical VLMs that can receive images as well as text as input, such as Flamingo and LLaVA, leverage existing LLMs (with frozen weights) and learn a mapping between the visual encoder and the LLM (Bordes et al. 2024). This approach is computationally more efficient than training both vision and text encoders from scratch, as it only requires learning the connection between modalities. To the extent that the LLM backbone achieves referential grounding from pre-training on text alone, our framework entails that such model should inherit this property. The addition of a superficial mapping to the embedding space of a visual encoder provides a form of sensorimotor grounding, but it does not contribute to the referential grounding inherited from the pre-trained LLM.

7.4. Embodiment

Embodiment is often mentioned as the best way to secure that AI systems are grounded in the world, via their robotic bodies. Some embodied systems like SayCan (Ahn et al. 2022) integrate a pre-trained LLM with a robotic body in a way that keeps these components functionally separate. In SayCan, an LLM breaks down verbal instructions into sequences of 'atomic' behaviours that the robot can enact through low-level visuomotor control. Importantly, the LLM component is not further trained or fine-tuned based on the success or failure of the robot's actions. In such systems, the embodied aspect adds nothing to the referential grounding of the language model component. Without additional training that would establish new causal-historical relations between the LLM's internal states and the outcomes of physical actions, the LLM's capacity for referential grounding remains exactly what it was before integration with the robotic system. Like multimodality, physical embodiment alone does not automatically contribute to referential grounding. What matters is the causal-historical relationship established through training that selects for internal states with world-involving functional roles. Without such a learning history, embodiment remains external to the representational system itself.

When fine-tuning is involved, the story is a bit different. Recent advances in robotic learning have led to the development of sophisticated multimodal systems that integrate vision, language, and action. Models such as RT-2 (Zitkovich et al. 2023), RT-X (O'Neill et al. 2024), and OpenVLA (Kim et al. 2024) typically build upon language models or vision-language models that were pre-trained on text or multimodal data before being fine-tuned for robotic control. These models incorporate a pre-trained language model component and may inherit referential grounding from it. For instance, RT-2 builds upon vision-language models like PaLI-X and PaLM-E, which themselves incorporate language models pre-trained on text data. To the extent that these underlying language models

achieve referential grounding through the mechanisms we have identified, the system as a whole may inherit this grounding.

However, VLA models go beyond merely inheriting referential grounding from their language model components. During embodied training, they are specifically trained to map visual inputs and natural language instructions to robotic actions that achieve specific goals in the physical world. This additional training process plausibly establishes new functional roles for the model's internal states that are directly tied to successful physical interactions.

From our perspective, embodied training potentially enhances referential grounding by establishing or reinforcing causal-historical relations between internal states and features of the world that are relevant for successful manipulation. The training process selects for internal states that carry information about these features precisely because doing so improves performance on the physical tasks. This introduces an additional source of normativity tied to success in physical interaction, complementing whatever referential grounding was inherited from the pre-trained components.

8. Concluding Remarks

We have argued that the Symbol Grounding Problem, as introduced by Harnad (1990), can be adapted to apply to deep neural networks such as Generative Pre-Trained Models in we have called the Vector Grounding Problem. Despite initial intuitions, we have sought to demonstrate that at least a subset of current Generative Pre-Trained Models have the features necessary to be endowed with representations and outputs grounded in the world.

While our discussion has focused on LLMs and, to a lesser extent, on their multimodal counterparts, any AI system fulfilling the causal-informational and historical relations requirements delineated above can possess representations with intrinsic meaning. Meeting these requirements is far from trivial, but nonetheless within reach of current AI models. As philosophical theories of representation have shown, in order to understand how cognitive systems – both biological and artificial – function, we must resist the temptation to assume that the presence of intrinsic meaning necessarily entails complex cognitive abilities such as understanding, knowledge, or consciousness. While having states imbued with intrinsic meaning might constitute one of the fundamental conditions for possessing these other cognitive abilities, it is insufficient on its own. Activations of specific populations of neurons in the mammalian cortex may represent the presence of edges in the visual scene, but these neuronal activations do not understand or possess knowledge about edges. We suggest that a similar claim can be made regarding the internal representations and outputs of LLMs and related connectionist models.

We argued for this claim in two steps. First, we distinguished between various forms of grounding, which are often conflated in the cognitive science and computer science literature. We suggested that one particular form of grounding, *referential grounding*, is central to both the Vector Grounding Problem and the older Symbol Grounding Problem. Referential grounding pertains to the way in which representations hook onto the world. It is the most fundamental type of grounding, and arguably serves as the basis for all other notions of grounding (with the partial exception of relational grounding). Referential grounding requires two main factors: causal-informational relations between internal states and the world, and historical relations that imbue those states with functions to carry information about the world.

Second, we have shown that certain LLMs trained exclusively on linguistic data, particularly those that undergo preference tuning, are referentially grounded. These models benefit from indirect causal-informational relations with worldly entities and properties, mediated by human interactions

with the world, which in turn shape the linguistic inputs they receive during training and inference. Preference tuning introduces an additional learning objective that is directly world-involving, as it encompasses a requirement for truthfulness or factuality. The normativity of this requirement supplies an extra-linguistic accuracy standard for the representations and outputs of LLMs. More tentatively, we have suggested that even without preference tuning, the representations and outputs of LLMs may achieve referential grounding in some limited domains, through pre-training and in-context learning on specific kinds of input sequences, such as those describing game moves.

This discussion led us to a counter-intuitive conclusion: some multimodal models trained on images in addition to text, like CLIP, are poor candidates for referential grounding, since they lack the kind of learning process that could supply world-involving functions. In fact, even some embodied systems that combine a 'frozen' LLM with a robotic body, like SayCan, also fail to acquire internal states and outputs that are referentially grounded. One benefit of our analysis is to highlight that multimodality and embodiment are neither necessary nor sufficient conditions for referential grounding in artificial systems, although they are certainly compatible with it if the right kind of learning process is available.

References

Ahn, M., Brohan, A., Brown, N., Chebotar, Y., Cortes, O., David, B., Finn, C., Fu, C., Gopalakrishnan, K., Hausman, K., Herzog, A., Ho, D., Hsu, J., Ibarz, J., Ichter, B., Irpan, A., Jang, E., Ruano, R. J., Jeffrey, K., Jesmonth, S., Joshi, N. J., Julian, R., Kalashnikov, D., Kuang, Y., Lee, K.-H., Levine, S., Lu, Y., Luu, L., Parada, C., Pastor, P., Quiambao, J., Rao, K., Rettinghouse, J., Reyes, D., Sermanet, P., Sievers, N., Tan, C., Toshev, A., Vanhoucke, V., Xia, F., Xiao, T., Xu, P., Xu, S., Yan, M. & Zeng, A. (2022), 'Do As I Can, Not As I Say: Grounding Language in Robotic Affordances'.

Artiga, M. (2016), 'New perspectives on artifactual and biological functions', *Applied Ontology* **11**, 89–102.

Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C. & Kaplan, J. (2021), 'A General Language Assistant as a Laboratory for Alignment'.

Barsalou, L. W. (1999), 'Perceptual symbol systems', *Behavioral and Brain Sciences* **22**, 577–660.

Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. (2021), On the dangers of stochastic parrots, *in* 'Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency', ACM.

Bender, E. M. & Koller, A. (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', Association for Computational Linguistics.

Block, N. (1986), 'Advertisement for a Semantics for Psychology', *Midwest Studies in Philosophy* **10**, 615–678.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D. E., Hong, J., Hsu, K., Huang, J., Icard, T., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., Keeling, G., Khani, F., Khattab, O., Koh, P. W., Krass, M., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X. L., Li, X., Ma, T., Malik, A., Manning, C. D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J. C., Nilforoshan, H., Nyarko, J., Ogut, G., Orr, L., Papadimitriou, I., Park, J. S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y., Ruiz, C., Ryan, J., Ré, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K., Tamkin, A., Taori, R., Thomas, A. W., Tramèr, F., Wang, R. E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S. M., Yasunaga, M., You, J., Zaharia, M., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K. & Liang, P. (2022), 'On the Opportunities and Risks of Foundation Models'.

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., Ibrahim, M., Hall, M., Xiong, Y., Lebensold, J., Ross, C., Jayakumar, S., Guo, C., Bouchacourt, D., Al-Tahan, H., Padthe, K., Sharma, V., Xu, H., Tan, X. E., Richards, M., Lavoie, S., Astolfi, P., Hemmat, R. A., Chen, J., Tirumala, K., Assouel, R., Moayeri, M., Talattof, A., Chaudhuri, K., Liu, Z., Chen, X., Garrido, Q., Ullrich, K., Agrawal, A., Saenko, K., Celikyilmaz, A. & Chandra, V. (2024), 'An Introduction to Vision-Language Modeling'.

Borg, E. (2025), 'LLMs, Turing tests and Chinese rooms: The prospects for meaning in large language models', *Inquiry* pp. 1–31.

Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., Driessche, G. B. V. D., Lespiau, J.-B., Damoc, B., Clark, A., Casas, D. D. L., Guy, A., Menick, J., Ring, R., Hennigan, T., Huang, S., Maggiore, L., Jones, C., Cassirer, A., Brock, A., Paganini, M., Irving, G., Vinyals, O., Osindero, S., Simonyan, K., Rae, J., Elsen, E. & Sifre, L. (2022), Improving Language Models by Retrieving from Trillions of Tokens, *in* 'Proceedings of the 39th International Conference on Machine Learning', PMLR, pp. 2206–2240.

Brandom, R. (1994), *Making It Explicit: Reasoning, Representing, and Discursive Commitment*, Harvard University Press.

Brennan, S. E. (1998), The grounding problem in conversations with and through computers, *in* S. R. Fussell & R. J. Kreuz, eds, 'Social and Cognitive Approaches to Interpersonal Communication', Lawrence Erlbaum, Hillsdale, NJ, pp. 201–225.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T. & Zhang, Y. (2023), 'Sparks of Artificial General Intelligence: Early experiments with GPT-4'.

Butlin, P. (2021), 'Sharing our concepts with machines', *Erkenntnis* **88**(7), 3079–3095.

Cappelen, H. & Dever, J. (2021), *Making AI Intelligible: Philosophical Foundations*, Oxford University Press, Oxford, New York.

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

Chalmers, D. J. (2025), 'Propositional Interpretability in Artificial Intelligence'.

Chandu, K. R., Bisk, Y. & Black, A. W. (2021), Grounding 'Grounding' in NLP, *in* 'Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021', Association for Computational Linguistics, Online, pp. 4283–4305.

Chomsky, N. (1957), *Syntactic Structures*, Mouton.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S. & Amodei, D. (2017), Deep Reinforcement Learning from Human Preferences, *in* 'Advances in Neural Information Processing Systems', Vol. 30, Curran Associates, Inc.

Clark, H. H. & Brennan, S. E. (1991), Grounding in communication, *in* 'Perspectives on Socially Shared Cognition', American Psychological Association, Washington, DC, US, pp. 127–149.

Coelho Mollo, D. (2015), 'Being clear on content', *Philosophia* **43**(3), 687–699.

Coelho Mollo, D. (2019), 'Are there teleological functions to compute?', *Philosophy of Science* **86**(3), 431–452.

Coelho Mollo, D. (2022), 'Deflationary Realism: Representation and idealisation in cognitive science', *Mind & Language* **37**(5), 1048–1066.

de Saussure, F. (1916), *Cours de Linguistique Générale*, Payot, Paris.

Delétang, G., Ruoss, A., Duquenne, P.-A., Catt, E., Genewein, T., Mattern, C., Grau-Moya, J., Wenliang, L. K., Aitchison, M., Orseau, L., Hutter, M. & Veness, J. (2024), 'Language Modeling Is Compression'.

Di Maro, M. (2021), 'Computational Grounding: An Overview of Common Ground Applications in Conversational Agents', *IJCoL. Italian Journal of Computational Linguistics* **7**(1 | 2), 133–156.

Firth, J. R. (1957), *Papers in Linguistics, 1934-1951*, Oxford University Press, London.

Fodor, J. A. (1990), *A Theory of Content and Other Essays*, MIT Press.

Garson, J. (2019), *What Biological Functions Are and Why They Matter*, Cambridge University Press.

Godfrey-Smith, P. (1994), 'A modern history theory of functions', *Noûs* **28**(3), 344–362.

Goldstein, S. & Levinstein, B. A. (2024), 'Does ChatGPT Have a Mind?'.

Grindrod, J. (2024), 'Large language models and linguistic intentionality', *Synthese* **204**(2), 71.

Grzankowski, A. (2024), 'Real sparks of artificial intelligence and the importance of inner interpretability', *Inquiry* **0**(0), 1–27.

Haiman, J. (1985), *Iconicity in Syntax*, Vol. 6, John Benjamins Publishing.

Harding, J. (2023), 'Operationalising Representation in Natural Language Processing'.

Harnad, S. (1990), 'The symbol-grounding problem', *Physica D* **42**, 335–346.

Harris, Z. S. (1954), 'Distributional structure', *Word* **10**, 146–162.

Heyes, C. M. (2018), *Cognitive gadgets*, The Belknap Press of Harvard University Press, Cambridge, Massachusetts. Includes bibliographical references and index.

Kim, M. J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P. & Finn, C. (2024), 'OpenVLA: An Open-Source Vision-Language-Action Model'.

LeCun, Y., Bengio, Y. & Hinton, G. (2015), 'Deep learning', *Nature* **521**(7553), 436–444.

Lederman, H. & Mahowald, K. (2024), 'Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of llms'.

Li, B. Z., Nye, M. & Andreas, J. (2021), Implicit Representations of Meaning in Neural Language Models, *in* 'Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)', Association for Computational Linguistics, Online, pp. 1813–1827.

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H. & Wattenberg, M. (2023), 'Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task'.

Mandelkern, M. & Linzen, T. (2024), 'Do Language Models' Words Refer?', *Computational Linguistics* pp. 1–10.

Mcclelland, J. L., Rumelhart, D. E. & Group, P. R. (1986), *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, MA.

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D. & Griffiths, T. L. (2024), 'Embers of autoregression show how large language models are shaped by the problem they are trained to solve', *Proceedings of the National Academy of Sciences* **121**(41).

Merrill, W., Wu, Z., Naka, N., Kim, Y. & Linzen, T. (2024), Can You Learn Semantics Through Next-Word Prediction? The Case of Entailment, *in* L.-W. Ku, A. Martins & V. Srikumar, eds, 'Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and Virtual Meeting, August 11-16, 2024', Association for Computational Linguistics, pp. 2752–2773.

Millière, R. (forthcoming), Language Models as Models of Language, *in* R. Nefdt, G. Dupre & K. Stanton, eds, 'The Oxford Handbook of the Philosophy of Linguistics', Oxford University Press.

Millière, R. & Buckner, C. (2024), 'A Philosophical Introduction to Language Models – Part I: Continuity With Classic Debates'.

Millière, R. & Buckner, C. (forthcoming), Interventionist Methods for Interpreting Deep Neural Networks, *in* G. Piccinini, ed., 'Neurocognitive Foundations of Mind', Oxford University Press, Oxford.

Millikan, R. G. (1989), 'In defense of proper functions', *Philosophy of Science* **56**(2), 288–302.

Millikan, R. G. (2017), *Beyond Concepts: Unicepts, Language, and Natural Information*, Oxford University Press.

Miracchi Titus, L. (2024), 'Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy', *Cognitive Systems Research* **83**, 101174.

Montague, R. (1970), 'Universal Grammar', *Theoria* **36**(3), 373–398.

Nanda, N., Lee, A. & Wattenberg, M. (2023), Emergent Linear Representations in World Models of Self-Supervised Sequence Models, *in* Y. Belinkov, S. Hao, J. Jumelet, N. Kim, A. McCarthy & H. Mohebbi, eds, 'Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2023, Singapore, December 7, 2023', Association for Computational Linguistics, pp. 16–30.

Neander, K. (1991), 'Functions as selected effects: The conceptual analyst's defence', *Philosophy of Science* **58**(2), 168–184.

Neander, K. (2017), *A Mark of the Mental: in defense of informational semantics*, The MIT Press.

O'Neill, A., Rehman, A., Gupta, A., Maddukuri, A., Gupta, A., Padalkar, A., Lee, A., Pooley, A., Gupta, A., Mandlekar, A., Jain, A., Tung, A., Bewley, A., Herzog, A., Irpan, A., Khazatsky, A., Rai, A., Gupta, A., Wang, A., Kolobov, A., Singh, A., Garg, A., Kembhavi, A., Xie, A., Brohan, A., Raffin, A., Sharma, A., Yavary, A., Jain, A., Balakrishna, A., Wahid, A., Burgess-Limerick, B., Kim, B., Schölkopf, B., Wulfe, B., Ichter, B., Lu, C., Xu, C., Le, C., Finn, C., Wang, C., Xu, C., Chi, C., Huang, C., Chan, C., Agia, C., Pan, C., Fu, C., Devin, C., Xu, D., Morton, D., Driess, D., Chen, D., Pathak, D., Shah, D., Büchler, D., Jayaraman, D., Kalashnikov, D., Sadigh, D., Johns, E., Foster, E., Liu, F., Ceola, F., Xia, F., Zhao, F., Frujeri, F. V., Stulp, F., Zhou, G., Sukhatme, G. S., Salhotra, G., Yan, G., Feng, G., Schiavi, G., Berseth, G., Kahn, G., Yang, G., Wang, G., Su, H., Fang, H.-S., Shi, H., Bao, H., Amor, H. B., Christensen, H. I., Furuta, H., Bharadhwaj, H., Walke, H., Fang, H., Ha, H., Mordatch, I., Radosavovic, I., Leal, I., Liang, J., Abou-Chakra, J., Kim, J., Drake, J., Peters, J., Schneider, J., Hsu, J., Vakil, J., Bohg, J., Bingham, J., Wu, J., Gao, J., Hu, J., Wu, J., Wu, J., Sun, J., Luo, J., Gu, J., Tan, J., Oh, J., Wu, J., Lu, J., Yang, J., Malik, J., Silvério, J., Hejna, J., Booher, J., Tompson, J., Yang, J., Salvador, J., Lim, J. J., Han, J., Wang, K., Rao, K., Pertsch, K., Hausman, K., Go, K., Gopalakrishnan, K., Goldberg, K., Byrne, K., Oslund, K., Kawaharazuka, K., Black, K., Lin, K., Zhang, K., Ehsani, K., Lekkala, K., Ellis, K., Rana, K., Srinivasan, K., Fang, K., Singh, K. P., Zeng, K.-H., Hatch, K., Hsu, K., Itti, L., Chen, L. Y., Pinto, L., Fei-Fei, L., Tan, L., Fan, L. J., Ott, L., Lee, L., Weihs, L., Chen,

M., Lepert, M., Memmel, M., Tomizuka, M., Itkina, M., Castro, M. G., Spero, M., Du, M., Ahn, M., Yip, M. C., Zhang, M., Ding, M., Heo, M., Srirama, M. K., Sharma, M., Kim, M. J., Kanazawa, N., Hansen, N., Heess, N., Joshi, N. J., Suenderhauf, N., Liu, N., Palo, N. D., Shafiullah, N. M. M., Mees, O., Kroemer, O., Bastani, O., Sanketi, P. R., Miller, P. T., Yin, P., Wohlhart, P., Xu, P., Fagan, P. D., Mitrano, P., Sermanet, P., Abbeel, P., Sundaresan, P., Chen, Q., Vuong, Q., Rafailov, R., Tian, R., Doshi, R., Mart'in-Mart'in, R., Baijal, R., Scalise, R., Hendrix, R., Lin, R., Qian, R., Zhang, R., Mendonca, R., Shah, R., Hoque, R., Julian, R., Bustamante, S., Kirmani, S., Levine, S., Lin, S., Moore, S., Bahl, S., Dass, S., Sonawani, S., Tulsiani, S., Song, S., Xu, S., Haldar, S., Karamcheti, S., Adebola, S., Guist, S., Nasiriany, S., Schaal, S., Welker, S., Tian, S., Ramamoorthy, S., Dasari, S., Belkhale, S., Park, S., Nair, S., Mirchandani, S., Osa, T., Gupta, T., Harada, T., Matsushima, T., Xiao, T., Kollar, T., Yu, T., Ding, T., Davchev, T., Zhao, T. Z., Armstrong, T., Darrell, T., Chung, T., Jain, V., Kumar, V., Vanhoucke, V., Zhan, W., Zhou, W., Burgard, W., Chen, X., Chen, X., Wang, X., Zhu, X., Geng, X., Liu, X., Liangwei, X., Li, X., Pang, Y., Lu, Y., Ma, Y. J., Kim, Y., Chebotar, Y., Zhou, Y., Zhu, Y., Wu, Y., Xu, Y., Wang, Y., Bisk, Y., Dou, Y., Cho, Y., Lee, Y., Cui, Y., Cao, Y., Wu, Y.-H., Tang, Y., Zhu, Y., Zhang, Y., Jiang, Y., Li, Y., Li, Y., Iwasawa, Y., Matsuo, Y., Ma, Z., Xu, Z., Cui, Z. J., Zhang, Z., Fu, Z. & Lin, Z. (2024), 'Open X-Embodiment: Robotic Learning Datasets and RT-X Models'.

OpenAI (2023), 'GPT-4 Technical Report'.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P. F., Leike, J. & Lowe, R. (2022), 'Training language models to follow instructions with human feedback', *Advances in Neural Information Processing Systems* **35**, 27730–27744.

Partee, B. H. (1981), Montague Grammar, Mental Representations, and Reality, *in* S. Kanger & S. Öhman, eds, 'Philosophy and Grammar: Papers on the Occasion of the Quincentennial of Uppsala University', Synthese Library, Springer Netherlands, Dordrecht, pp. 59–78.

Patel, R. & Pavlick, E. (2022), Mapping Language Models to Grounded Conceptual Spaces, *in* 'International Conference on Learning Representations'.

Pavlick, E. (2022), 'Semantic structure in deep learning', *Annual Review of Linguistics* **8**(1), 447–471.

Pepp, J. (2025), 'Reference without intentions in large language models', *Inquiry* pp. 1–19.

Perniss, P. & Vigliocco, G. (2014), 'The bridge of iconicity: From a world of experience to the experience of language', *Philosophical Transactions of the Royal Society B: Biological Sciences* **369**(1651), 20130300.

Petroni, F., Rocktäschel, T., Lewis, P., Bakhtin, A., Wu, Y., Miller, A. H. & Riedel, S. (2019), 'Language Models as Knowledge Bases?'.

Piantadosi, S. T. & Hill, F. (2022), 'Meaning without reference in large language models'.

Poon, H. (2013), Grounded Unsupervised Semantic Parsing, *in* 'Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)', Association for Computational Linguistics, Sofia, Bulgaria, pp. 933–943.

Pulvermüller, F. (1999), 'Words in the brain's language', *Behavioural and Brain Sciences* **22**, 253–336.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G. & Sutskever, I. (2021), 'Learning Transferable Visual Models From Natural Language Supervision', *arXiv:2103.00020 [cs]* .

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D. & Finn, C. (2024), 'Direct Preference Optimization: Your Language Model is Secretly a Reward Model'.

Searle, J. R. (1980), 'Minds, brains and programs', *Behavioral and Brain Sciences* **3**(3), 417–457.

Shea, N. (2018), *Representation in Cognitive Science*, Oxford University Press.

Shortliffe, E. H. (1977), 'Mycin: A Knowledge-Based Computer Program Applied to Infectious Diseases', *Proceedings of the Annual Symposium on Computer Application in Medical Care* pp. 66–69.

Simon, H. A. & Newell, A. (1971), 'Human problem solving: The state of the theory in 1970', *American Psychologist* **26**(2), 145–159.

Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., Brown, A. R., Santoro, A., Gupta, A., Garriga-Alonso, A., Kluska, A., Lewkowycz, A., Agarwal, A., Power, A., Ray, A., Warstadt, A., Kocurek, A. W., Safaya, A., Tazarv, A., Xiang, A., Parrish, A., Nie, A., Hussain, A., Askell, A., Dsouza, A., Slone, A., Rahane, A., Iyer, A. S., Andreassen, A. J., Madotto, A., Santilli, A., Stuhlmüller, A., Dai, A. M., La, A., Lampinen, A., Zou, A., Jiang, A., Chen, A., Vuong, A., Gupta, A., Gottardi, A., Norelli, A., Venkatesh, A., Gholamidavoodi, A., Tabassum, A., Menezes, A., Kirubarajan, A., Mullokandov, A., Sabharwal, A., Herrick, A., Efrat, A., Erdem, A., Karakaş, A., Roberts, B. R., Loe, B. S., Zoph, B., Bojanowski, B., Özyurt, B., Hedayatnia, B., Neyshabur, B., Inden, B., Stein, B., Ekmekci, B., Lin, B. Y., Howald, B., Orinion, B., Diao, C., Dour, C., Stinson, C., Argueta, C., Ferri, C., Singh, C., Rathkopf, C., Meng, C., Baral, C., Wu, C., Callison-Burch, C., Waites, C., Voigt, C., Manning, C. D., Potts, C., Ramirez, C., Rivera, C. E., Siro, C., Raffel, C., Ashcraft, C., Garbacea, C., Sileo, D., Garrette, D., Hendrycks, D., Kilman, D., Roth, D., Freeman, C. D., Khashabi, D., Levy, D., González, D. M., Perszyk, D., Hernandez, D., Chen, D., Ippolito, D., Gilboa, D., Dohan, D., Drakard, D., Jurgens, D., Datta, D., Ganguli, D., Emelin, D., Kleyko, D., Yuret, D., Chen, D., Tam, D., Hupkes, D., Misra, D., Buzan, D., Mollo, D. C., Yang, D., Lee, D.-H., Schrader, D., Shutova, E., Cubuk, E. D., Segal, E., Hagerman, E., Barnes, E., Donoway, E., Pavlick, E., Rodolà, E., Lam, E., Chu, E., Tang, E., Erdem, E., Chang, E., Chi, E. A., Dyer, E., Jerzak, E., Kim, E., Manyasi, E. E., Zheltonozhskii, E., Xia, F., Siar, F., Martínez-Plumed, F., Happé, F., Chollet, F., Rong, F., Mishra, G., Winata, G. I., de Melo, G., Kruszewski, G., Parascandolo, G., Mariani, G., Wang, G. X., Jaimovitch-Lopez, G., Betz, G., Gur-Ari, G., Galijasevic, H., Kim, H., Rashkin, H., Hajishirzi, H., Mehta, H., Bogar, H., Shevlin, H. F. A., Schuetze, H., Yakura, H., Zhang, H., Wong, H. M., Ng, I., Noble, I., Jumelet, J., Geissinger, J., Kernion, J., Hilton, J., Lee, J., Fisac, J. F., Simon, J. B., Koppel, J., Zheng, J., Zou, J., Kocon, J., Thompson, J., Wingfield, J., Kaplan, J., Radom, J., Sohl-Dickstein, J., Phang, J., Wei, J., Yosinski, J., Novikova, J., Bosscher, J., Marsh, J., Kim, J., Taal, J., Engel, J., Alabi, J., Xu, J., Song, J., Tang, J., Waweru, J., Burden, J., Miller, J., Balis, J. U., Batchelder, J., Berant, J., Frohberg, J., Rozen, J., Hernandez-Orallo, J., Boudeman, J., Guerr, J., Jones, J., Tenenbaum, J. B., Rule, J. S., Chua, J., Kanclerz, K., Livescu, K., Krauth, K., Gopalakrishnan, K., Ignatyeva, K., Markert, K., Dhole, K., Gimpel, K., Omondi, K., Mathewson, K. W., Chiafullo, K., Shkaruta, K., Shridhar, K., McDonell, K., Richardson, K., Reynolds, L., Gao, L., Zhang, L., Dugan, L., Qin, L., Contreras-Ochando, L., Morency, L.-P., Moschella, L., Lam, L., Noble, L., Schmidt, L., He, L., Oliveros-Colón, L., Metz, L., Senel, L. K., Bosma, M., Sap, M., Hoeve, M. T., Farooqi, M., Faruqui, M., Mazeika, M., Baturan, M., Marelli, M., Maru, M., Ramirez-Quintana, M. J., Tolkiehn, M., Giulianelli, M., Lewis, M., Potthast, M., Leavitt, M. L., Hagen, M., Schubert, M., Baitemirova, M. O., Arnaud, M., McElrath, M., Yee, M. A., Cohen, M., Gu, M., Ivanitskiy, M., Starritt, M., Strube, M., Swędrowski, M., Bevilacqua, M., Yasunaga, M., Kale, M., Cain, M., Xu, M., Suzgun, M., Walker, M., Tiwari, M., Bansal, M., Aminnaseri, M., Geva, M., Gheini, M., T, M. V., Peng, N., Chi, N. A., Lee, N., Krakover, N. G.-A., Cameron, N., Roberts, N., Doiron, N., Martinez, N., Nangia, N., Deckers, N., Muennighoff, N., Keskar, N. S., Iyer, N. S., Constant, N., Fiedel, N., Wen, N., Zhang, O., Agha, O., Elbaghdadi, O., Levy, O., Evans, O., Casares,

P. A. M., Doshi, P., Fung, P., Liang, P. P., Vicol, P., Alipoormolabashi, P., Liao, P., Liang, P., Chang, P. W., Eckersley, P., Htut, P. M., Hwang, P., Miłkowski, P., Patil, P., Pezeshkpour, P., Oli, P., Mei, Q., Lyu, Q., Chen, Q., Banjade, R., Rudolph, R. E., Gabriel, R., Habacker, R., Risco, R., Millière, R., Garg, R., Barnes, R., Saurous, R. A., Arakawa, R., Raymaekers, R., Frank, R., Sikand, R., Novak, R., Sitelew, R., Bras, R. L., Liu, R., Jacobs, R., Zhang, R., Salakhutdinov, R., Chi, R. A., Lee, S. R., Stovall, R., Teehan, R., Yang, R., Singh, S., Mohammad, S. M., Anand, S., Dillavou, S., Shleifer, S., Wiseman, S., Gruetter, S., Bowman, S. R., Schoenholz, S. S., Han, S., Kwatra, S., Rous, S. A., Ghazarian, S., Ghosh, S., Casey, S., Bischoff, S., Gehrmann, S., Schuster, S., Sadeghi, S., Hamdan, S., Zhou, S., Srivastava, S., Shi, S., Singh, S., Asaadi, S., Gu, S. S., Pachchigar, S., Toshniwal, S., Upadhyay, S., Debnath, S. S., Shakeri, S., Thormeyer, S., Melzi, S., Reddy, S., Makini, S. P., Lee, S.-H., Torene, S., Hatwar, S., Dehaene, S., Divic, S., Ermon, S., Biderman, S., Lin, S., Prasad, S., Piantadosi, S., Shieber, S., Misherghi, S., Kiritchenko, S., Mishra, S., Linzen, T., Schuster, T., Li, T., Yu, T., Ali, T., Hashimoto, T., Wu, T.-L., Desbordes, T., Rothschild, T., Phan, T., Wang, T., Nkinyili, T., Schick, T., Kornev, T., Tunduny, T., Gerstenberg, T., Chang, T., Neeraj, T., Khot, T., Shultz, T., Shaham, U., Misra, V., Demberg, V., Nyamai, V., Raunak, V., Ramasesh, V. V., vinay uday prabhu, Padmakumar, V., Srikumar, V., Fedus, W., Saunders, W., Zhang, W., Vossen, W., Ren, X., Tong, X., Zhao, X., Wu, X., Shen, X., Yaghoobzadeh, Y., Lakretz, Y., Song, Y., Bahri, Y., Choi, Y., Yang, Y., Hao, Y., Chen, Y., Belinkov, Y., Hou, Y., Hou, Y., Bai, Y., Seid, Z., Zhao, Z., Wang, Z., Wang, Z. J., Wang, Z. & Wu, Z. (2023), 'Beyond the imitation game: Quantifying and extrapolating the capabilities of language models', *Transactions on Machine Learning Research* .
URL: *https://openreview.net/forum?id=uyTL5Bvosj*

Stalnaker, R. (2002), 'Common Ground', *Linguistics and Philosophy* **25**(5-6), 701–721.

Sterelny, K. (2014), *The evolved apprentice*, The Jean Nicod lec tures, 1st paperback edition edn, The MIT Press, Cambridge, Massachusetts.

Stoljar, D. & Zhang, Z. V. (2025), 'Why ChatGPT doesn't think: An argument from rationality', *Inquiry* pp. 1–29.

Thoppilan, R., De Freitas, D., Hall, J., Shazeer, N., Kulshreshtha, A., Cheng, H.-T., Jin, A., Bos, T., Baker, L., Du, Y., Li, Y., Lee, H., Zheng, H. S., Ghafouri, A., Menegali, M., Huang, Y., Krikun, M., Lepikhin, D., Qin, J., Chen, D., Xu, Y., Chen, Z., Roberts, A., Bosma, M., Zhao, V., Zhou, Y., Chang, C.-C., Krivokon, I., Rusch, W., Pickett, M., Srinivasan, P., Man, L., Meier-Hellstern, K., Morris, M. R., Doshi, T., Santos, R. D., Duke, T., Soraker, J., Zevenbergen, B., Prabhakaran, V., Diaz, M., Hutchinson, B., Olson, K., Molina, A., Hoffman-John, E., Lee, J., Aroyo, L., Rajakumar, R., Butryna, A., Lamm, M., Kuzmina, V., Fenton, J., Cohen, A., Bernstein, R., Kurzweil, R., Aguera-Arcas, B., Cui, C., Croak, M., Chi, E. & Le, Q. (2022), 'LaMDA: Language Models for Dialog Applications'.

Tishby, N. & Zaslavsky, N. (2015), 'Deep Learning and the Information Bottleneck Principle'.

Tomasello, M. (2009), *Cultural Origins of Human Cognition*, Harvard University Press, Cambridge. Description based upon print version of record.

Traum, D. R. (1994), A Computational Theory of Grounding in Natural Language Conversation, Ph.d. thesis, University of Rochester, Rochester, NY.

Tsai, C.-T. & Roth, D. (2016), 'Concept Grounding to Multiple Knowledge Bases via Indirect Supervision', *Transactions of the Association for Computational Linguistics* **4**(0), 141–154.

Van Langendonck, W. (2010), Iconicity, *in* D. Geeraerts & H. Cuyckens, eds, 'The Oxford Handbook of Cognitive Linguistics', Oxford University Press, p. 0.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017), 'Attention is all you need'.

von Oswald, J., Schlegel, M., Meulemans, A., Kobayashi, S., Niklasson, E., Zucchet, N., Scherrer, N., Miller, N., Sandler, M., y Arcas, B. A., Vladymyrov, M., Pascanu, R. & Sacramento, J. (2024), 'Uncovering mesa-optimization algorithms in Transformers'.

Wolfram, S. (2023), 'What is chatgpt doing ... and why does it work?', BlogPost, accessed on 23.03.2023.
URL: *https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/*

Wright, L. (1973), 'Functions', *The Philosophical Review* **82**(2), 139–168.

Wu, X. & Varshney, L. R. (2024), A Meta-Learning Perspective on Transformers for Causal Language Modeling, *in* L.-W. Ku, A. Martins & V. Srikumar, eds, 'Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and Virtual Meeting, August 11-16, 2024', Association for Computational Linguistics, pp. 15612–15622.

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D. & Zou, J. (2023), 'When and why vision-language models behave like bags-of-words, and what to do about it?'.

Zitkovich, B., Yu, T., Xu, S., Xu, P., Xiao, T., Xia, F., Wu, J., Wohlhart, P., Welker, S., Wahid, A., Vuong, Q., Vanhoucke, V., Tran, H. T., Soricut, R., Singh, A., Singh, J., Sermanet, P., Sanketi, P. R., Salazar, G., Ryoo, M. S., Reymann, K., Rao, K., Pertsch, K., Mordatch, I., Michalewski, H., Lu, Y., Levine, S., Lee, L., Lee, T.-W. E., Leal, I., Kuang, Y., Kalashnikov, D., Julian, R., Joshi, N. J., Irpan, A., Ichter, B., Hsu, J., Herzog, A., Hausman, K., Gopalakrishnan, K., Fu, C., Florence, P., Finn, C., Dubey, K. A., Driess, D., Ding, T., Choromanski, K. M., Chen, X., Chebotar, Y., Carbajal, J., Brown, N., Brohan, A., Arenas, M. G. & Han, K. (2023), RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control, *in* J. Tan, M. Toussaint & K. Darvish, eds, 'Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA', Vol. 229 of *Proceedings of Machine Learning Research*, PMLR, pp. 2165–2183.