What it's like to be an LLM

Ryan M. Nefdt[1,2,*]

[1]Department of Philosophy, University of Cape Town, South Africa
[2]Department of Philosophy, University of Bristol, UK
[*]contact: ryan.nefdt@uct.ac.za/ryan.nefdt@bristol.ac.uk

Abstract

This article is not about machine consciousness. It's about our understanding of new technology. An emerging contemporary trend treats large language models (LLMs) as cognitive beings partly because they display high-level linguistic abilities. This is in part because we struggle to conceive of a purely linguistic agent. I flesh out this new typological possibility while suggesting that there are interesting features attributable to LLMs based on their architectures and the kinds of information they can process that help us to see the world through their nodes.

Keywords: *Cognition*, *LLMs*, *Philosophy of AI*, *linguistic agency*

1. Introduction

In his famous article, Nagel (1974) makes a compelling case for the inaccessibility of conscious experience such as those of non-human animals. His argument is a challenge to reductionism and the idea that objective (or third personal) facts can give us insight into first person or perspectival experiences. Ultimately, the contested concept of consciousness plays an important role in the debates that followed (Dennett, 1991a; Hacker, 2002; Schwitzgebel, 2000). Consciousness is not my interest here, neither human, machine, nor chiropteran. Rather, I explore the cognitive consequences of a new kind of hypothetical creature, a purely and completely linguistic being, the possibility of which is exemplified by the current suite of large language models (LLMs).

This is as much an issue in the philosophies of mind and language as it is in the philosophy of technology. A prominent trend in both academia and everyday folk psychology is to treat LLMs as if they are cognitive agents or incorporate underlying cognitive processes such as beliefs, desires, and understanding. Part of the reason for this tendency is that we have never had to grapple with a linguistic non-cognitive entity or system before in the wild. Mahowald et al. (2024) call this the 'good at language → good at thought' fallacy. They state it in the following way:

> The popularity of the Turing test, combined with language–thought coupling in everyday life, has led to several common fallacies related to the language–thought relationship. One fallacy is that an entity (be it a human or a machine) that is good at language must also be good at thinking. If an entity generates coherent stretches of text, it must possess rich knowledge and reasoning capacities. (Mahowald et al., 2024, p. 517)

Language has always been a conduit to understanding or accessing cognition in humans in some form or other (long before Turing) from the Sapir-Whorf hypothesis (see Pelletier & Nefdt, 2025) to the language of thought hypothesis (Fodor, 1975).[1]

In the philosophy of AI, the position I advocate occupies a mid-point between the emerging trend of motivating full machine cognition (Cappelen & Dever, 2025) and thorough eliminativism about LLMs (Bender & Koller, 2020a). In §3, I identify a taxonomy of views on the LLM cognitive spectrum and I place my general position within a novel quadrant. The rest of the paper attempts to draw out the philosophical consequences of this view, i.e. what the world might look like to an LLM. In §4.1, I argue that experience is not required for perspective, which I suggest LLMs can possess. In §4.2, I describe the complicated relationship between LLMs and time. Here, I make the case that different architectures might produce different answers to the titular question. Lastly, in §4, I qualify the position generated by these arguments and suggest one way to appreciate what it might be like to be an LLM.

2. What is an LLM?

Let's say you wanted to designed an artificial system that could process and produce natural language. You could approach your task in a number of different ways. You could try the 'good old fashioned' way. Program a set of rules that are triggered by a certain kind of input to generate systematic output. The method is meant to mirror how people are said to represent syntax and semantics. It's a compositional mereological procedure. In a sense, you'd be treating a computer like a linguist or an expert system. The rules could be those that linguists have honed in their studies of various natural language structures, like if you have a noun and a determiner then you have a noun phrase (or a determiner phrase depending on your preferred theory, see Pullum & Miller, 2022). This is a uniquely symbolic approach. It also starts at a mature or adult linguistic competence. It involves using symbols to represent objects and rules to manipulate those objects via the symbols. Programming languages still retain this kind of explicit instruction procedure. The approach has a certain logical flair to it. It's clear and transparent (even though anyone who has written thousands of lines of code might disagree). In this sense, a 'large' language model could be a (very large) data or knowledge base with hand crafted rules for extracting patterns and generating results. But this is not what the term is usually taken to mean today.

Modern large language models are a marriage of connectionist architectures, Big Data, and machine learning algorithms designed to recognise and replicate patterns in the data. In a sense it starts with the perspective of treating the process like a child's learning of language. Modern variants such as Gemini, ChatGPT, or DeepSeek are multilayered networks of nodes or 'artificial neurons' connected by weights or parameters numbering in the billions. They are trained on massive corpora and quite probably the entire internet. Artificial neural nets are a strictly broader class of machine learning systems than language models, that includes image classifiers like convolutional networks and generative adversarial networks etc. But our focus will be on those marshalled in the use of natural language processing.

Like most machine learning systems, LLMs go through two important phases: training and testing. During training, LLMs are fed tokenised text in the form of vectors. Initially, they randomly guess the connections between the vectors. But after some back and forth

[1] It is telling that many developmental psychological tests, such as the Weschsler Intelligence Scale for Children or the MacArthur-Bates CDI use linguistic skills as a primary measure of cognitive development.

between output and the desired target, they incrementally adjust their parameters to produce accurate output (or local optima). This output usually takes the form of a probability distribution over the vocabulary towards the goal of generating the next likely token in a string. So if the LLM is computing 'the cat is on the' it will produce 'mat' based on what it has seen in the data (or some missing or masked word like 'cat' if its a cloze task, also used in training). This task of next token production is called 'autoregression' and it basically translates any problem into a problem of predictive processing (see McCoy et al. (2024) for some limitations of this process). Take what you've seen before and guess the most likely next word. But the 'back-propagation algorithm' (the aforementioned back and forth error minimisation technique) only applies to training. During the testing phase, LLMs are strictly feed-forward. Here the idea is to test the models on unseen data and see if they can generalise, i.e. use what they 'learned' during training to novel input. The chatbots we interact with on a daily basis are in the feed-forward mode with an additional component of having undergone some reinforcement learning with human feedback (RLHF), where their outputs are further scrutinised for accuracy, potential harm, and usefulness.

So LLMs take text as input and generate text as output. As simple as this sounds, it has led to surprisingly diverse applications from text summarisation, and question answering to translation.[2] Some philosophers have speculated that this is due to the fact that natural language is the most efficient compression of thought (Rothschild, forthcoming). In other words, process language and you get thought for free. I'll dispute this claim in what follows.

Nevertheless, given the linguistic medium in which these AI systems operate, it is easy to slip into anthropomorphic terminology when describing the behaviour of LLMs. This is partly because they seem to 'talk back' and other AI systems don't generally do that. In fact, take the hyperparameter of temperature as an example. It is a control setting which allows variation in the randomness of LLM output. What this means is that at lower temperatures, the system will stick to the script (find the most likely next word) but at higher temperature settings they will adjust the probability distribution to make riskier choices resulting in more 'human-like' creative sequences of sentences. Going forward, we won't presuppose a definitive answer on how to individuate LLMs (see Jones et al. (2026)). We will also return to some specifics about their architecture in §4.2 and when that makes a difference to our central exploration. For now, this functional level of individuation should suffice (for more technical details see Millière (2024) and Millière & Buckner (2024)).

3. What it means to be purely linguistic

3.1 The Missing Quadrant

It is an undeniable, yet contingent, fact that cognition and language are deeply connected. Linguistic development in humans is linked to cognitive development so much so that it is often a measurement of it (see footnote 1). On the flip-side, impaired cognition often results in impaired linguistic ability (but not always the other way around, see Fedorenko et al., 2024b). Prominent theories of linguistics have described the remit of the theorising as that of cognitive science (Chomsky, 1965, 1968; Jackendoff, 2002; Langacker, 2008). In other words, studying language has been taken as studying the mind or brain of language

[2]See Balashov (2025) for a comparison between LLMs and bespoke machine translation systems.

users.[3]

Thus, it seems uncontroversial to describe humans as cognitive creatures endowed with language, i.e. cognitive and linguistic agents. But cognition abounds across species even if language doesn't follow. The ability to process environmental information and/or represent that information internally or collectively stretches from primates to octopus and ant colonies. This makes it reasonable to state that there are creatures with cognition or cognitive abilities which lack language.[4]

	Language	Cognition
Humans	X	X
Non-human animals	-	X
???	X	-

Table 1: Conceptual possibilities

As we can see in table 1, cognition and language have found a home in our species so much so that this amalgam has prompted many scholars (from Darwin to Dennett) to attribute our particular niche or advantage to its union. But they are separable both conceptually and in practice, as the wealth of evidence from non-human biology attests. Non-human animals display a wide range of cognitive abilities without language.[5]

The central question of this paper is then whether there exists creatures which are linguistic but not cognitive. In other words, are there any systems that occupy the missing quadrant of the table? I will argue in the affirmative. Thus, LLMs occupy a hitherto vacant part of conceptual space. In a way, they are a proof of concept. Depending on your favoured theory of mind, LLMs would then count as exemplifying mindless (or disembodied) language. Chalmers (2023) has a slightly different take in which a 'disembodied thinker' like an LLM could in principle have a form of cognitive consciousness.

You might worry that linguistic machines have been around for decades. ELIZA was pioneered in the 1960s. It was designed to pattern match and simulate human conversation. It was a symbolic program, with specific 'scripts' written in a LISP-like language. Despite its limitations, humans interacting with this first chatbot began to attribute feelings and emotions to it (a phenomenon known as the 'Eliza effect', Natale, 2021). However, ELIZA was not a fully and complete linguistic agent. The program could only operate on a limited script and used a keyword substitution technique that produced often shallow and incoherent language. It was not generative in the sense of modern LLMs, vastly impoverished in terms of its capabilities, and could not 'learn' natural language in any deep sense. It had little to no reasonable sense of autonomy (we return to agency and autonomy below).

The contemporary debate in philosophy can be interpreted as a struggle over the same conceptual real estate. Cappelen & Dever (2025) make the bold argument that LLMs are in line with humans in terms of both cognition and language.

ChatGPT is a full-blown linguistic and cognitive agent—on par with humans.

[3]This approach to understanding language has even prompted some to deny thought to non-linguistic creatures (Davidson, 1975, 2001). Language and thought are a package deal. In fact, thought itself is often analysed as thinking in a language (Fodor, 1975). But I will take cognition to be broader than 'representational' thought here, largely obviating these thorny issues.

[4]As a example of the extent of this cognitive reach, the field of basal cognition provides many case studies (Lyon et al., 2021).

[5]You could, of course, insist that those also possessing communicative signalling systems do indeed possess language. But we will sidestep that issue here and stipulate that they lack a full-blown language as that found in our species.

> It can use language meaningfully, making assertions, asking and answering questions, offering suggestions, and giving commands. Its use of language reflects underlying mental states—it can know and believe things, desire things, and wonder about things. It can learn about the world (by being told about it, or reasoning from what it already knows). It can make plans and reason about how to achieve them. It can take actions to implement those plans. (Cappelen & Dever, 2025, p. 13)

I think their arguments do establish some sense of (linguistic) agency (more below) but do not support the claim of more substantive cognition. They have a number of subarguments for their position. One powerful line is what they call the 'Just an X fallacy', which Millière & Buckner (2024, pp. 9–10) call the 're-description fallacy'.

> This fallacy arises when critics argue that a system cannot model a particular cognitive capacity, simply because its operations can be explained in less abstract and more deflationary terms. In the present context, the fallacy manifests in claims that LLMs could not possibly be good models of some cognitive capacity ϕ because their operations merely consist in a collection of statistical calculations, or linear algebra operations, or next-token predictions.

In a sense, I think this is too strong. The strategy I take will indeed 're-describe' the cognitive capacity attributions of LLMs but not necessarily in deflationary terms. The burden will be on advocates of stronger cognitive claims to motivate their positions beyond just the behavioural output of LLMs. Agüera y Arcas (2022, 2025b) similarly argues that artificial intelligence has been achieved by LLMs. Furthermore, he insists that scaling computation was actually the mechanism that led to this situation. His view is complex and involves evolutionary biology, artificial life, cybernetics, computer science, and automata theory (among others). He argues that modern AI is evidence of a kind of 'technological symbiogenesis' in which multiple parallel computational elements conjoined to produce intelligence. Support comes from running simulations on computer programs like Bff and witnessing sudden complexity emerging from simple initial strings of code (see Agüera y Arcas (2025a) for more details). The important point is that he thinks these processes significantly resemble biological evolutionary processes in both degree and kind, specifically the move from non-living systems (without inherited traits) to living ones (abiogenesis). Needless to say, if he considers LLMs to be truly intelligent then he attributes a kind of cognition, albeit computational, to them. Cappelen & Dever (2025) offer a more mentalistic and synchronic take on machine cognition and intelligence. Their main positive argument involves what they call the 'holistic network assumption' in which mental and intentional features are deeply interconnected and self-reinforcing. For instance, any system that displays *understanding* appreciates *meanings* and any system that can *answer* questions displays *knowledge* which in turn requires *beliefs* and other mental states. Lastly, actions imply intentions and goals reasoning. They make the case that LLMs such as ChatGPT embody this interconnected network of behaviours and features. I have alternative explanations for their assertions (and to a certain extent those of Aguera y Arcas).[6]

My overarching alternative does not involve the eliminativism associated with Bender & Koller (2020b) and others or the re-description fallacy *tout court*. It simply aims to etch

[6] I have lumped Aguera y Arcas and Cappelen & Dever together here as examples of the fully cognitive and linguistic position. But these projects are significantly distinct. The latter see themselves as providing an epistemic, structural argument, opponents of which should counter. Whereas the former is engaged in direct theorising about the capacities of AI.

out a more minimal position on machine cognition, namely the idea that LLMs are purely and fully linguistic agents. This view, I hold, has the resources to explain the network of Cappelen & Dever (2025) and the computational evolution story of Agüera y Arcas (2025a) to some extent.

3.2 Cognition Unplugged

What does it mean to occupy the missing quadrant of table 1? What does it mean to be a purely linguistic agent or system? The mere logical possibility doesn't provide us with any direct clues. For instance, is understanding or knowledge or belief precluded from a purely linguistic interaction with the world?

The clues actually come from a number of different angles including recent neuroscience and the limitations of current AI models. Let us start with the latter. In human beings, language processing is one part of a larger cognitive whole. Some linguists have distinguished between the faculty of language narrowly construed and the faculty broadly construed (Hauser et al., 2002). Basically, the purely linguistic part (often associated with syntactic competence) is responsible for processing and producing linguistic forms while the broader structure incorporates semantic and other cognitive processes. In LLMs, it's more complicated than this. Syntax and semantics (and pragmatics) are not easily separated in the data.[7]. As Potts (2019) notes concerning deep learning approaches to NLP: "[learned vectors] will reflect many aspects of language use: biases in word frequency, preferences for certain readings, pragmatic refinements of lexical items, and so forth" (e118). Let us ignore (but not forget) this complication for the moment. The point is that modern NLP only models one part of the larger cognitive system associated with language while neglecting other non-linguistic information processing in humans.

In a recent interview, influential AI expert Andrej Karpathy makes a similar observation when he claims that engineers have probably recreated a kind of cortical tissue in terms of pattern-learning, "but we're still missing the rest of the brain. No hippocampus for memory. No amygdala for instincts. No emotions or motivations."[8] This brings me to my first qualification on cognition of LLMs, they only isolate or filter a particular kind of linguistic information and neglect other brain regions.[9]

> NO BRAINER: LLMs only model one aspect of cognition, namely (statistical) linguistic processing.

One could argue that that they don't even really do this (as eliminativists are want to do) since they process tokenised text by reducing that to numerical vectors and computations on that vector space. But this may or may not be a case of the redescription fallacy. Despite the pervasiveness of language (and the fraught connections between language and thought), the restrictions that NO BRAINER induces are palpable.[10] It rules out claims of higher cognitive capabilities in LLMs. But this doesn't mean that certain

[7] There are also many linguists who reject the autonomy of syntax and the competence-performance distinctions (Goldberg & Suttle, 2010)

[8] https://www.youtube.com/watch?v=lXUZvyajciY

[9] This doesn't apply to AI in general as their are interesting deep learning approaches to vision (arguably ones that ushered in the epoch), audio processing, and robotics. But these models are not yet multimodal and thus are not incorporated within LLMs.

[10] Consider research projects like Lindsey (2025) which aim to measure the extent to which models like Claude Opus 4.1 can be said to be 'introspective'. As interesting as this possibility might be, to turn pure linguistic processing into metacognitive capacity requires modelling more parts of the brain than just collocational patterns in language.

facsimiles of cognitive abilities are not present within exclusive linguistic processing, for this qualification we turn to neuroscience.

Here we draw on recent work in neurolinguistics to guide our philosophical discussion. Casto et al. (2025) make a distinction that helps to fill in our concept of 'pure linguistic' agency. Their position is described as follows:

> [W]e argue that a **deep understanding** of language, which entails building mental models and rich representations of meaning that connect to our broader knowledge of the world, requires the exportation of information from the brain's core language system to other cognitive and neural systems that can build models of what we are hearing or reading. (Casto et al., 2025, p. 2)

To understand the position, they claim that we need to appreciate what is known as a 'language network' operating in the brain, revealed by fMRI and impairment studies (Fedorenko et al., 2024a). I don't think this is necessary (however compelling the evidence). Other modular views in linguistics and cognitive science that isolate linguistic knowledge or processing would do just as well, such as Pylyshyn's notion of 'cognitive impenetrability' of a module (Pylyshyn, 1980). All we need is a reasonable separation between the processing of language (or linguistic input) and that of other cognitive modalities such as vision, reasoning, planning, motor cognition etc.[11] For some neuroscientists, the evidence shows specific selection for language in particular networks. With this in place, we can qualify different intentional states with the 'linguistic' modifier in a nontrivial way. Following Casto et al. (2025), *linguistic* understanding differs from 'deep understanding' in nature. Importantly, they make this distinction in humans. They hypothesise that there is a (shallow) level of understanding language that the language systems trades in. It involves knowledge of statistical co-occurrence and structural information between linguistic objects. Impairment to these regions can leave cognitive and conceptual processing largely intact. Thus, the language system produces representations of language but lacks a deeper appreciation of its role in the world and larger cognitive structures. Again, studies on non-linguistic animals further support the conceptual ablation of language from cognition.

Casto et al. (2025) are not blind to the connections with this picture and that of LLMs and AI. They compare the core language system's ability, to not understand language deeply but only appreciate how people talk about language, to text-based LLMs. They argue that these models display formal competence in producing linguistic forms (patterns and rules) but not functional competence which requires understanding and usage in the world (mirroring the argument in Mahowald et al. (2024)). The basic point is that language can be conceptually unplugged from the rest of cognition and this process can involve some appreciation of the relationships between words and concepts at a shallow level. LLMs are the proof of concept here. This can explain the Holistic Network Assumption of intentions posited by Cappelen & Dever (2025). Answering questions, understanding prompts, acting on these prompts and knowing the anwers all come with the *linguistic* (or shallow) modifier.

But even with this conceptual ablation in place, a further step is needed to reach the competencies of LLMs. LLMs are not just analogous to the core linguistic system but a vastly amplified version of it. Here, the scaling point of Agüera y Arcas (2025b) comes to the fore.[12] What is interesting is that scaling this subset of human cognition has been so effective in approximating other systems such as reasoning. This might have

[11] What's interesting here is that we don't even need consensus in cognitive science. The fact that divergent views converge on the isolation of the language system is illustrative.

[12] To address the arguments of Agüera y Arcas (2025a) is more complicated as he draws on a number

something to do with how language compresses reasoning (Rothschild, forthcoming) or built-in redundancies in the cognitive system as a whole. Either way, LLMs involve unplugged intentional states.[13]

> Cognition Unplugged: Pure linguistic agency has statistically-based proxies for more cognitively loaded states. But they are unplugged from the rest of cognition and the world.

Earlier simpler programs like ELIZA do not yet approach or approximate the core language system as modern LLMs do. But neither do they involve agency in any reasonable sense. Before making some positive claims about what it might be like to be a purely linguistic agent or LLM, let me say a word about what I mean by 'agency'.

3.3 Agency and Autonomy

The term 'AI agents' is commonplace in the technology industry and increasing in its usage in the public space. They are considered autonomous software systems that can perform complex tasks which little to no prompting. They are bought and shared in the online marketplace. But in what way are they agents in a more philosophically-loaded sense?

Opinions differ in this arena. Hopster & Löhr (2023) highlight the idea of LLM agency as a potential case of 'conceptual misalignment'. They criticise the language of agency used by the Future of Life Institute by arguing (alongside the DAIR institute) that accountability is misattributed to artifacts and not developers in this case. Thus those who promote such claims "misconstrue the concept of agency, which is suggestive of various other capacities and responsibilities that cannot be properly ascribed to LLMs" (Hopster & Löhr, 2023, p. 70). Their worry is about the normative dimension of agency and in this sense, it is quite warranted. But the behavioural, psychological level of agency is independent of the larger normative issues. Most (if not all) animals are not moral agents. Thus, responsibility for their actions (such as attacks on humans) cannot easily be attributed to them. But they are agents in the sense of autonomously engaging with their environments. This is the sense in which software agents are *agents*.

In the philosophy of technology, outside the realm of generative AI, similar arguments have been made. Consider the specific example of social media algorithms, such as those found in TikTok, Instagram, and YouTube. Such algorithms are tasked with continuously suggesting video content for users based on their preferences, history, and possible future interests and of course they do not do so randomly.[14] They are carefully designed to achieve their goals. The overall goal of the engineers and executives, who determine the algorithms, is to drive engagement in one way or another for its users (the further goal may be maximising shareholder revenue). The action guidance part is expressed in terms of engagement numbers. The actions of the algorithm succeed when the numbers go up. The AI in this case is the Intelligent Software Agents (ISA) with which an individual user interacts. Burr et al. (2018) draw on AI, behavioural economics, control theory, and game

of fields. My sense is that armed with *linguistic* understanding suitably qualified and synthetic biology, we can similarly explain artificial intelligence (specifically LLMs) by analogy with non-living systems like viruses and collective intelligence in biology, see more in §4.

[13] You might counter that LLMs are plugged into the world via the internet. Indeed retrieval augmented generation (RAG) does provide a mechanism for extracting information from online sources. Still this interaction is purely linguistic and pattern-based.

[14] For a description of the algorithm from the company itself, see https://www.youtube.com/howyoutubeworks/recommendations/

theory to describe the behaviour of ISAs such as YouTube's algorithm. They emphasise important parallels between the action guiding behaviour of humans and ISAs in the sense that ISAs are driven by goals, can make autonomous decisions, and learn from experience.

> The simplest models assume an agent trying to pick the best action in a given state, where "best" refers to whatever maximises expected utility or reward. The agent can (partially) observe the current state of the environment, compute the probability for each of the possible outcomes of each action, and then evaluate each outcome on the basis of its expected utility. (Burr et al., 2018, p. 738)

This, for them, constitutes a type of 'bounded rationality'. This framework further allows for the characterisation of how these algorithms or agents *coerce* and even *deceive* users within complex feedback loops (cf. AI nudging Calboli & Engelen (2025)).

LLMs are agents in this goal-driven, partially autonomous, and learning sense of the concept. There are perhaps more loaded concepts involving responsibility and normative embedding that they lack but ironically enforcing such restrictions seems to ascend to a level of anthropomorphism, albeit in the opposite direction.

4. The World through Language

In the previous section, a largely negative case was presented. There I was interested in carving out a novel possibility in conceptual space. In this last section, I attempt to say something positive about what it might be like to be an LLM. However, I will qualify this position in section §4.3. Let me begin with two properties of cognition I do think are reasonably present in LLMs, perspective and time. Although, the latter introduces further complications at a more fine-grained level of analysis.

4.1 Perspective without experience

Certain views on consciousness link the concept to perspective and experience. What it's like to be a bat or a sperm whale is then essentially tied up in the world from the perspective of that creature. I've argued that LLMs can have language or be equipped with it without the corresponding cognition we assume of humans. In this section, I will argue that they can have perspective without experience.

Let us start with Nagel's classic connection between conscious experience and perspective. He argues that reductionist (third personal or scientific views) are unlikely to capture what it is like for an animal to experience the world.

> If physicalism is to be defended, the phenomenological features must themselves be given a physical account. But when we examine their subjective character it: seems that such a result is impossible. The reason is that every subjective phenomenon is essentially connected with a single point of view, and it seems inevitable that an objective, physical theory will abandon that point of view. (Nagel, 1974, p. 437)

What is interesting here is that LLMs can be trained to execute a particular point of view. However, this point of view is purely linguistic and is not necessarily accompanied by phenomenal experience or even deep understanding. As an example, consider the Dennett-bot created by Schwitzgebel et al. (2024). In this research, the authors created

a language model fine-tuned on the work of the late philosopher and cognitive scientist Daniel Dennett. In a kind of Turing test, they then asked philosophers and lay participants alike to distinguish Dennett's real answers to a set of philosophical questions from that of the bot. Ordinary people performed at near chance levels while philosophers (trained in Dennett's work) did so above chance but still below expectation. Dennett himself claimed that the bot answered some questions in a more Dennettian manner than he himself did. What can we learn from this anecdotal evidence?

One thing is that perspective too has a proxy in purely linguistic agents. They can be trained to filter their behaviour through the information gained in a specific corpus. On a slightly more superficial level, prompts can include this kind of code switching instructions as evidenced by students who use LLMs to write their assignments or people who use them as therapists.[15] More integrated processes might attach experience to perspective (i.e. moving from the core language network to other cognitive systems) but filtering language with language does not by itself imply this further concept.

Another problem is at the level of individuation. I think to further complicate Nagel's intuitions, we have to ask whose perspective is at stake. LLMs are distributed information processing systems. Their user-interface (or 'instance agents', Goldstein & Lederman, 2025) is ephemeral and can result in different outcomes based on individual prompt environments. Birch (2025) describes our individualisation of chatbots in terms of a 'persistent interlocutor illusion'. To me, this suggests that LLMs are more closely aligned to the perspective of eusocial insects or labs of individual researchers. A lab could have a perspective in the sense of favouring a specific theoretical framework but asking what it's like to be a lab or an ant colony is different from asking what it's like to be an individual researcher or ant. Perhaps it would be more appropriate to ask what it's like to be *in* an LLM. However, I think that the individual phenomenal level is missing in the case of deep learning systems like LLMs. Again, perspective can still be described but it is not directly analogous to the human individual experiencer case.

4.2 Data Structures and Time

What about time? Do or could LLMs possess a concept of time in any quantifiable sense? This is complicated for a number of reasons and can only receive a qualified answer. For that answer we need to get into the weeds to a certain extent. Consider Agüera y Arcas (2022, p. 189) on this issue.

> One fundamental difference between large language models like GPT-3 or LaMDA and biological brains is that brains operate continuously in time. For language models, time as such does not really exist, only conversational turns in strict alternation, like moves in a game of chess. Within a conversational turn, letters or words are emitted sequentially with each "turn of the crank."

It is not clear what "time as such" means here. In the philosophy of time, the experience of time is often considered indistinguishable from the experience of things changing. Certainly the perception of the passage of time is measured in terms of the rate of changing events (Le Poidevin, 2007; Shoemaker, 1993). Turn taking should suffice to produce this effect, even if long pauses and resets are part of the process. Perhaps the disconnected and multiple parallel sessions make a single linear sequence of events impossible to attribute to an LLM. Metaphysicians have speculated on the possibility of 'disunified'

[15] Nagel's argument might also be confronted with the wealth of work on observer effects in the natural sciences. Perspective is not unapproachable by scientific theorising. See also Dennett (1991a).

time series and largely concluded that it is incoherent to imagine unconnected and parallel time series (Le Poidevin, 2007; Quinton, 1993). It does indeed seem unlikely that something could experience such as temporal sequence outside of science fiction.[16]

The case I am making for LLM temporality is slightly distinct from these broader claims. It relies in part on a distinction made in Klein (2025) between *sublinear*, *linear*, and *supralinear* formats of representation. Before that a bit of background.

Until very recently a number of different architectures were used to model different kinds of cognitive tasks. In computer vision, Convolutional Neural Networks (CNNs) were the norm (Buckner, 2019; LeCun et al., 2010). The central mechanism of CNNs is convolution.[17] This is a process that creates a layer in the network that is sensitive to specific feature and transforms an image accordingly. The convolution layer applies a 'kernel' or filter to the input image to extract various aspects of it. For instance, one such feature extraction operation is 'max pooling', which "reduces the spatial dimensions of features by selecting the maximum value within each small window or region" (Zhao & Zhang, 2024, p. 5). In a sense, the convolution process creates a kind of surface image or 'visual buffer' (by creating a smaller image with different dimensions).

In NLP, Recurrent Neural Networks (RNNs) were standardly used for sequence-to-sequence translation and other linguistic tasks. However, AI changed dramatically with the advent of Transformer technology (Vaswani et al., 2017). Suddenly, there was a general purpose model that could outperform CNNs on image recognition, RNNs on NLP tasks, and various other architectures in the domain for which they were designed.[18] At the heart of a transformer is a self-attention mechanism. Self-attention, as the name suggests, allows the model to selectively focus on different parts of the input sequence instead of treating everything the same way, or contextualising with only the previous sequence as with RNNs. Tranformers are also more efficient to train and incorporate much larger context windows.

Returning to Klein (2025), he worries that transformers have the wrong kind of representational format to be useful models for the cognitive science of language. His three-way distinction favours what he calls 'supralinear' formats like graphs or arrays which are multiply realisable in terms of linear formats, i.e. structured objects in which elements can be arranged in various ways.[19] He claims that

> The level of representational family is also a useful one for cognitive science, because many classic debates about the format of cognition really come down to debates about the representational family needed to achieve certain competences. The conclusion is nearly always that something supralinear is required. (Klein, 2025, p. 5)

Transformers for Klein embody two kinds of invariance (permutation and substring) which together limit the kinds of formats they can represent to purely linear ones. But

[16] Pelletier & Nefdt (2025) make an interesting analogy between the Heptapod species of aliens in the short story by Ted Chiang and LLMs in terms of linguistic relativity.

[17] In mathematics, a convolution is an integral that expresses the amount of overlap of one function as it is shifted over another function.

[18] This impression has somewhat dampened over time with multiple mixed models and even a resurgence of interest in older architectures such as RNNs, see Feng et al. (2024).

[19] This is a handy way to interpret the arguments of Moro et al. (2023) who claim that LLMs process impossible languages in the same way they do possible ones, thereby rendering them useless for linguistics. Possible languages in their case means (supralinear) hierarchical graph-like structures like syntactic trees. Again, a distinction emerges in terms of architecture as the original experiments were run on RNNs (Mitchell & Bowers, 2020) while the follow up work, with different conclusions, involved transformers (Kallini et al., 2024).

humans, he argues, are unlikely to process language that way. Basically the idea is that transformers don't care about the order of elements or position of tokens when they store information, hence the need for positional encoding. They don't have structured representations in this way. In RNNs, on the other hand, position and order matters since "for language modelling [they] present elements sequentially and use the effect of earlier computations to alter processing of later ones, trying to build a format with sensitivity to syntactic structure" (Klein, 2025, p. 11). This might explain why RNNs emerged from the more cognitive-inspired connectionist approaches to linguistic representation. Similarly CNNs have structured formats of representations for vision. Klein thinks that LLMs are therefore poor models of how we represent language but good models of corpora, which he thinks is a nontrivial achievement.

What does this mean for time? Well, it could mean that different architectures encode temporal sequences differently or in a more cognitively realistic manner. In other words, they appreciate time in language to varying degrees (starting from zero). Transformers might have an implicit block universe with atemporal or timeless picture of information while RNNs might take a tensed view in which order matters. The fact that transformers have proven to be useful across domains is a further clue as to their generality (i.e. unstructured formats of representation). But domains like language imply time in important ways and the less efficient LLMs might actually track these properties more faithfully. Of course, future architectures might process temporal order in other ways hitherto undiscovered.

This isn't to say that LLMs do or do not have a sense of time. But rather to say that they could depending on the kinds of structures they employ in processing natural language information.

4.3 Why I'm neither a realist nor an eliminativist

Dennett (1991b) motivates what he considers to be a 'semi-realist' position on mindedness, based on the concept of real patterns in information theory. Interestingly, real pattern analysis has been harnessed with relation to the connection between formal linguistics and LLMs (Futrell & Mahowald, 2025; Nefdt, 2023). Instrumentalist approaches modelled on the intentional stance have also been used to explain LLM intentionality (Lederman & Mahowald, 2024). Dennett's position is pitched between the thoroughgoing realist and the instrumentalist cum eliminativist. It is unclear whether he succeeded in carving this niche. But the ideal position lies somewhere between taking minds seriously and dismisses or reducing them to quantifiable properties. Similarly, I think we can resist both the whole hog realism of AI cognitive agency put forward by Cappelen & Dever (2025) and the eliminativist conclusions of Bender & Koller (2020b) and the like. There is a nonempty position in between in which LLMs are purely linguistic agents unplugged from integration with both larger cognitive structure and the world in which it evolved.

In other words, there is something it's like to be an LLM, equipped with linguistic understanding, perspective and even possibly a sense of time (depending on the model) but this something is not an experience of the world as such and at this stage of the technology does not entail cognitive intentional states in the ways we usually assume they exist in humans and some non-human animals. I'll leave it up to the reader to decide whether enough has been said to motivate this cautious optimism. To some extent it evokes an image of a 'world in the data' (Ladyman & Ross, 2013) as opposed to a world model built from it.

5. Conclusion

In this article, I have attempted to address the question of what it is like to be an LLM. Specifically, I have proffered a novel position in the conceptual landscape on LLM cognition. I have argued that there is a sense in which LLMs are purely linguistic agents. This possibility necessarily limits the potential for deeper notions of cognition to apply to their workings. I have marshalled arguments from neuroscience and AI to show that this does not mean that LLMs can't understand *simpliciter* and they might even possess a kind of collective perspective and a sense of time, all without an accompanying deeper cognition or intentionality.

Acknowledgements

I have so many people to thank for their useful and critical comments on this research. I am grateful to James Ladyman, Max Jones, and Emmanuele Ratti at the University of Bristol for various discussions on the content. Herman Cappelen, Rachel Sterken and the audience at the AI & Humanity Lab at the University of Hong Kong provided both necessary challenges and insights into how I might fulfil my lofty aims. Audiences at Waseda University in Tokyo were also instrumental in the final draft presented here. And lastly, Jeroen Hopster and researchers at the institute for the Ethics of Socially Disruptive Technologies at Utrecht added further layers of consideration and engagement to the work. I blame everyone of these people for all the footnotes as well as the lack of sweeping conclusions!

References

Agüera y Arcas, B. (2022). Do large language models understand us? *Daedalus*, 151(2), 183–197. https://doi.org/10.1162/daed_a_01909

Agüera y Arcas, B. (2025a). *What is intelligence?: Lessons from AI about evolution, computing, and minds*. The MIT Press. https://mitpress.mit.edu/9780262049955/what-is-intelligence/

Agüera y Arcas, B. (2025b). What is the future of intelligence? the answer could lie in the story of its evolution. *Nature*, 647, 846–850. https://doi.org/10.1038/d41586-025-03857-0

Balashov, Y. (2025). Translation in the wild. *Information*, 16(12), 1077. https://doi.org/10.3390/info16121077

Bender, E. M., & Koller, A. (2020a). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (Eds.), *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bender, E. M., & Koller, A. (2020b). Climbing towards NLU: On meaning, form, and understanding in the age of data. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Birch, J. (2025). Ai consciousness: A centrist manifesto. *PsyArXiv*. https://doi.org/10.31234/osf.io/af7c9_v1

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14(10), e12625. https://doi.org/10.1111/phc3.12625

Burr, C., Cristianini, N., & Ladyman, J. (2018). An analysis of the interaction between intelligent software agents and human users. *Minds and Machines*, 28(4), 735–774. https://doi.org/10.1007/s11023-018-9479-0

Calboli, S., & Engelen, B. (2025). Ai-enhanced nudging in public policy: Why to worry and how to respond. *Mind & Society*, 24, 529–547. https://doi.org/10.1007/s11299-025-00322-3

Cappelen, H., & Dever, J. (2025). Going whole hog: A philosophical defense of ai cognition. https://arxiv.org/abs/2504.13988

Casto, C., Ivanova, A., Fedorenko, E., & Kanwisher, N. (2025). What does it mean to understand language? https://arxiv.org/abs/2511.19757

Chalmers, D. J. (2023). Could a large language model be conscious? *Boston Review*, 1.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N. (1968). *Language and mind*. Harper; Row.

Davidson, D. (1975). Thought and talk. In S. D. Guttenplan (Ed.), *Mind and language* (pp. 1975–7). Clarendon Press.

Davidson, D. (2001). What thought requires. In J. Branquinho (Ed.), *The foundations of cognitive science* (p. 121). Oxford University Press UK.

Dennett, D. C. (1991a). *Consciousness explained*. Penguin Books.

Dennett, D. C. (1991b). Real patterns. *The Journal of Philosophy*, 88(1), 27–51. https://doi.org/10.2307/2027085

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024a). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25, 289–312.

Fedorenko, E., Ivanova, A. A., & Regev, T. I. (2024b). The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5), 289–312. https://doi.org/10.1038/s41583-024-00802-4

Feng, L., Tung, F., Ahmed, M. O., Bengio, Y., & Hajimirsadegh, H. (2024). Were rnns all we needed? https://arxiv.org/abs/2410.01201

Fodor, J. (1975). *The language of thought*. Harvard University Press.

Futrell, R., & Mahowald, K. (2025). How linguistics learned to stop worrying and love the language models [Published online 24 July 2025]. *Behavioral and Brain Sciences*, 1–98. https://doi.org/10.1017/S0140525X2510112X

Goldberg, A. E., & Suttle, L. (2010). Construction grammar. *WIREs Cognitive Science*, 1(4), 468–477. https://doi.org/10.1002/wcs.22

Goldstein, S., & Lederman, H. (2025). *What does chatgpt want? an interpretationist guide* [https://philarchive.org/rec/GOLWDC-2].

Hacker, P. M. S. (2002). Is there anything it is like to be a bat? *Philosophy*, 77(300), 157–174. https://doi.org/10.1017/s0031819102000220

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science*, 298(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Hopster, J., & Löhr, G. (2023). Conceptual engineering and philosophy of technology: Amelioration or adaptation? *Philosophy & Technology*, 36, 70. https://doi.org/10.1007/s13347-022-00548-7

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press.

Jones, M., Ladyman, J., & Nefdt, R. M. (2026). *Counting (on) large language models* [https://philpapers.org/rec/JONCOL].

Kallini, J., Papadimitriou, I., Futrell, R., Mahowald, K., & Potts, C. (2024). Mission: Impossible language models. https://arxiv.org/abs/2401.06416

Klein, C. (2025). What do language models model? transformers, automata, and the format of thought. https://arxiv.org/abs/2508.18598

Ladyman, J., & Ross, D. (2013). The world in the data. In D. Ross, J. Ladyman & H. Kincaid (Eds.), *Scientific metaphysics*. Oxford University Press.

Langacker, R. (2008, February). *Cognitive grammar: A basic introduction*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195331967.001.0001

Le Poidevin, R. (2007). *Images of time: An essay on temporal representation*. Oxford University Press.

LeCun, Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, 253–256. https://doi.org/10.1109/ISCAS.2010.5537907

Lederman, H., & Mahowald, K. (2024). Are language models more like libraries or like librarians? bibliotechnism, the novel reference problem, and the attitudes of llms. *Transactions of the Association for Computational Linguistics*, 12, 1087–1103. https://doi.org/10.1162/tacl_a_00690

Lindsey, J. (2025). Emergent introspective awareness in large language models [Published October 29, 2025]. *Transformer Circuits Thread*. https://transformer-circuits.pub/2025/introspection/index.html

Lyon, P., Keijzer, F., Arendt, D., & Levin, M. (2021). Reframing cognition: Getting down to biological basics. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1820), 20190750. https://doi.org/10.1098/rstb.2019.0750

Mahowald, K., Ivanova, A. A., Blank, I. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2024). Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6), 517–540. https://doi.org/10.1016/j.tics.2024.01.011

McCoy, R. T., Yao, S., Friedman, D., Hardy, M. D., & Griffiths, T. L. (2024). When a language model is optimized for reasoning, does it still show embers of autoregression? an analysis of OpenAI o1. https://arxiv.org/abs/2410.01792

Millière, R. (2024). Language models as models of language. https://arxiv.org/abs/2408.07144

Millière, R., & Buckner, C. (2024). A philosophical introduction to language models – part I: Continuity with classic debates. https://arxiv.org/abs/2401.03910

Mitchell, J., & Bowers, J. (2020, December). Priorless recurrent networks learn curiously. In D. Scott, N. Bel & C. Zong (Eds.), *Proceedings of the 28th international conference on computational linguistics* (pp. 5147–5158). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.451

Moro, A., Greco, M., & Cappa, S. F. (2023). Large languages, impossible languages and human brains. *Cortex*, 167, 82–85. https://doi.org/https://doi.org/10.1016/j.cortex.2023.07.003

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83(4), 435–50. https://doi.org/10.2307/2183914

Natale, S. (2021). The ELIZA effect: Joseph weizenbaum and the emergence of chatbots. In S. Natale (Ed.), *Deceitful media: Artificial intelligence and social life after the turing test* (pp. 73–96). Oxford University Press. https://doi.org/10.1093/oso/9780190080365.003.0004

Nefdt, R. M. (2023). *Language, science, and structure: A journey into the philosophy of linguistics*. Oxford University Press.

Pelletier, F. J., & Nefdt, R. (2025). *Linguistic relativity: An essential guide to past debates and future prospects*. Oxford University Press.

Potts, C. (2019). A case for deep learning in semantics: Response to Pater. *Language*, 95(1), e115–e124. https://doi.org/10.1353/lan.2019.0019

Pullum, G. K., & Miller, P. (2022). *NPs versus DPs: Why chomsky was right* [LingBuzz archive, paper no. 6845]. https://lingbuzz.net/lingbuzz/006845

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, 3(1), 111–169.

Quinton, A. (1993). Space and times. In R. Le Poidevin & M. MacBeath (Eds.), *The philosophy of time* (pp. 203–220). Oxford University Press.

Rothschild, D. (forthcoming). Language and thought: The view from llms. In D. Sosa & E. Lepore (Eds.), *Oxford studies in philosophy of language volume 3*. Oxford University Press.

Schwitzgebel, E. (2000). How well do we know our own conscious experience? the case of human echolocation. *Philosophical Topics*, 28(2), 235–46. https://doi.org/10.5840/philtopics20002824

Schwitzgebel, E., Schwitzgebel, D., & Strasser, A. (2024). Creating a large language model of a philosopher. *Mind & Language*, 39(2), 237–259.

Shoemaker, S. (1993). Time without change. In R. Le Poidevin & M. MacBeath (Eds.), *The philosophy of time* (pp. –). Oxford University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://arxiv.org/abs/1706.03762

Zhao, L., & Zhang, Z. (2024). An improved pooling method for convolutional neural networks. *Scientific Reports*, 14, 1589. https://doi.org/10.1038/s41598-024-04289-7