What do LLMs tell us about phenomenology?*

Jeff Yoshimi

Abstract: Large language models are in many ways neurally and psychologically unrealistic. They are disembodied, quasi-omniscient, uniform in tone, and implemented in a circuit which bears little resemblance to the human brain. I argue that they nonetheless provide a plausible explanation for a specific feature of the phenomenology of language: our capacity to draw effortlessly on unconscious background knowledge while hearing and speaking. I associate this ability with "meaning-saturated appropriate-access circuits": non-propositional structures which store and rapidly retrieve relevant information. In the course of the discussion, I draw on Dreyfus' critique of AI, Husserl's account of sedimentation, and Mangan's account of rightness as a coherence signal in the "fringe" of peripheral consciousness.

Keywords: Husserl, Neural Networks, Large Language Models, Dreyfus, Mangan

1. Introduction

The answer to the titular question is: not much, but not nothing either. And that "not nothing" is worth taking seriously, or so I shall argue.

My approach will be something like a compare and contrast while zooming. First, I'll outline the basics of large language models (or LLMs), the form of AI that runs familiar chatbots like ChatGPT. Then, I'll consider broadly how LLMs are similar to and different from humans. I'll then focus on a specific set of features of LLMs parallel to the phenomenology of language use, and then I'll immediately consider differences. In doing so I will isolate a specific circuit in LLMs that sheds light on and may explain the phenomenology of language use and unconscious access.

One might object: "Who cares? Any two objects can be compared. Fish and accordions are very different, but alike in some ways. Unlike fish, accordions can't be eaten and don't run or

* Penultimate draft of a paper to appear in Ferrarello, Susi (ed.), *Phenomenology, Cognition and Artificial Intelligence*, Springer.

swim, but *like* fish, they are physical objects made up of molecules and taking up space." In response it can be noted that the comparison has some limited use: fish and accordions both have a "space-efficient surface expansion system". Like the gills of a fish, the bellows of an accordion can fold up to save space, but also expand to expose surface area. That is a common design principle between the two.

Something similar is going on with LLMs and phenomenology. One feature of human phenomenology is the ability to use language. From the vast reservoirs of past experience we pull out the right words needed in any situation. LLMs provide a plausible explanation of this ability, using what I will call "meaning-saturated appropriate-access circuits".[1] These circuits are used in a way that is quite different from how corresponding circuits are used in the human brain. Nonetheless I will argue that they are a key piece of the puzzle of how neural networks in the brain give rise to phenomenology.

I am working in the tradition of naturalized phenomenology or neurophenomenology (Gallagher 2012; Varela 1996; Yoshimi 2016). Note that Husserl and other phenomenologists are open to interdisciplinary work of this kind, so long as extra-phenomenological insights are also confirmed in reflection (Yoshimi 2016; Yoshimi 2022).[2]

[1] I use "meaning" in a phenomenological sense: when someone sees a dog or talks about a cat, their experiences of seeing and talking have meanings associated with dogs and cats, respectively. A "circuit" in the brain related to these meanings produces them when activated. A circuit in a computer model can simulate such a biological circuit. A circuit is "meaning-saturated" or "super-saturated" when it supports many more experiences of meaning than its size would suggest.

[2] A paper with similar aims to this, focused on the concept of synthesis, is (Fazi 2024).

2. Connectionist Preliminaries

A large language model or LLM is a type of feed-forward neural network typically implemented using the transformer architecture. Most language models are built using the transformer architecture, so I will take "LLM" to be shorthand for "transformer-based large language model."[3] LLMs are the dominant technology in current generative AI systems, including familiar chat models like ChatGPT.

LLMs are neural networks so we start with a brief history of the field. There is a long tradition of understanding the dynamics of mind and thought in terms of flowing activation in networks in the brain (Yoshimi et al. 2025). The laws of association developed by the empiricists were often linked to the strengthening of connections in the brain. Freud's unfinished project for a scientific psychology linked the psychodynamics of mental phenomena with neuro-dynamic processes in the brain (Freud, 1966). In the 20th century these ideas were formalized into mathematical models, and by the 1970s it was possible to build and train networks to perform simple classification tasks like recognizing different shapes in a retinal array.

Since that time most production-level neural networks have been built using the same broad method: wire up the neurons in a feed-forward way (with no recurrent connections), and train them to perform a task using a calculus-based method called "backpropagation" or simply "backprop"[4]. All connections in such a network flow forward from input layer to hidden layers to output layer via weighted connections. You tell the network to do something using a table that

[3] Language models can be small and can be built using other architectures, and the transformer architecture can be used for tasks besides language modeling, but in practice most language models are large and are built using the transformer architecture.

[4] Backprop can be adapted to certain kinds of recurrent networks, but it is primarily used for feed-forward networks.

associates inputs with targets, and backprop incrementally adjusts the weights in such a way that error (the difference between actual and target values) is reduced. Networks are initialized with random weights, and so behavior is initially poor; but as the weights are changed the intended behavior slowly manifests.

A prominent early example was reading text aloud in a child's voice (Sejnowski and Rosenberg 1987). At first the network makes random sounds when it "reads", but then it slowly, and somewhat creepily, learns to pronounce regular phonemes and eventually produce coherent speech. When networks like this first emerged in the 1980s it created a huge wave of interest in the emerging field of cognitive science. These "connectionist" networks seemed to be learning in the same way humans do. Across many domains of human behavior, from vision to language to sensorimotor coordination, neural network models learned to behave in remarkably realistic ways (Rumelhart and McClelland 1986; Doerig et al. 2023).

Part of the appeal of these networks was how different they were from traditional symbolic AI models, which were hand-programmed to do things like recognize patterns and answer questions. Schank and Abelson's classic story-understanding model illustrates the symbolic approach. It contained a set of rules like "If asked what is on the menu, read this script" (Schank and Abelson 1977). The system could answer questions about, for example, the process of ordering a meal in a restaurant. These systems often became quite complicated, using sophisticated rule systems and databases to approximate human performance. The neural network approach was quite different. Rather than being hand-programmed, networks were trained using the backprop procedure described above, where weights are incrementally changed in a way that reduces error.

This basic difference generates a range of philosophically important differences that we will need to keep in mind while thinking about LLMs. In symbolic AI, it is more or less known how a given system works. The way a model responds to a question can be traced to specific rules and database queries. In a neural network, by contrast, we have a giant mess of neurons and connections, and it's often not possible to clearly describe how they do something. Backprop feels like magic. We say what we want, the math is applied, and it delivers. Some people refer to this as "growing" rather than designing networks.

It's worth pausing to think about how strange this is. We build systems we don't fully understand. That is rare and perhaps unprecedented in the history of engineering. We grow networks to produce coherent speech, but it's not (currently) possible to say exactly *how* they do this.[5] However, tools have emerged for studying these networks, especially in the field of mechanistic interpretability, also referred to as "mechinterp" or "interpretability" research (Olah et al. 2020; Elhage et al. 2021). The field is reminiscent of cognitive neuroscience. Probes are built to detect when certain kinds of things are happening: at what layer is the network processing syntax or semantics, or recognizing textures? The patterns are hard to find, because they are holistically entangled with many other patterns. But mathematical techniques have emerged for understanding these patterns as directions in high-dimensional spaces (Beckmann and Queloz 2026). Once these directions are found, they can be used to perform "ablations", where a given function is turned off (a bit like a lesion in the brain), or they can be "steered", enhancing the function. For example, a chat model might have its humor function ablated, so that it stops using humor, or enhanced, so that it uses humor all the time. Features that have been

[5] Paul Humphreys has described a similar phenomenon with "epistemic opacity" (Humphreys 2009). This is a radical and morally-fraught form of opacity: even the people who built it don't understand it.

studied in this way include part-of-speech, sentiment, geographic location, and numerical structure (Tenney, Das, and Pavlick 2019; Tigges et al. 2024; Gurnee and Tegmark 2024; Zhou et al. 2024).[6]

Another noteworthy feature of neural networks is that they are "realistic without being realistic." They are in many ways quite different from real brains, especially in the way they are trained using backprop. And yet, once they are trained, they behave a lot like real brains. As David Zipser (1992) asks: "Why should model neurons, particularly the 'hidden units' which are not directly instructed what to do, develop response properties that so closely resemble real neurons?" He notes that this is especially surprising "because most of the algorithms used to train the networks are not biologically plausible." Still, the right task demands applied to an unrealistic circuit can still produce realistic behaviors.[7] I will call this "Zipser's dictum."

3. The Transformer Architecture

The connectionist revolution of the 1980s and 90s produced many compelling models, but the approach fizzled out in the early 2000s, as limitations of small feed-forward networks trained on relatively small datasets emerged. In the 2010s researchers figured out how to effectively train "deep" networks on much larger datasets. These networks had many hidden layers, sometimes over a hundred. It was in this period that neural networks started being used regularly by large companies like Google, especially for image and speech recognition. As

[6] A helpful website which allows for a quick sense of how *millions* of features are encoded in LLMs is www.neuronpedia.org.

[7] The situation is perhaps most striking in vision. For years people tried to build custom models of the visual system based on how the brain works. But these bespoke models were beat out by neural networks (originating in industrial applications) that were simply trained to recognize objects (Yamins et al. 2014).

interest in neural networks grew, researchers got better at using backprop in a modular way. By the late 2010s, researchers had several toolboxes available which they could use to wire up feed-forward networks however they wanted, and "grow" them with backprop. It was in this period that LLMs first emerged.

A crucial step in this development was the introduction of the transformer architecture. In 2017, researchers at Google built a special kind of layer that could learn to pay attention to information flowing through the network, grabbing and processing relevant data (Vaswani et al. 2017). When they trained this type of model on language tasks, it worked quite well, and it was soon realized these models could be wired up as chat bots.

Figure 1 gives a basic sense of how LLMs work. The inputs and outputs are, as with any feed-forward network, sets of numbers. On the left is a context window. This can be thought of as a giant stack of tokens (a token can be a word, a part of a word, a bit of punctuation, or an emoji). Everything in a standard chat, including what you say, what the AI says back (and various hidden things as well) is contained in the context window. To adumbrate a bit, it helps to think of the context window as a kind of working memory system, containing all the information that the network "has in mind" at a time. When the network is updated, each token in the stack is converted into a list of numbers (a vector) and those numbers are passed through all the layers and weights of the neural network. The final layer, the output layer, has a single neuron or "node" for each possible token in the network's vocabulary (about 50-250K in production models) and it outputs a probability distribution over these. At any iteration, the token corresponding to one of the most activated output nodes is selected, and that token is added back to the context window. Then the process repeats. In this way a chat takes place. In Figure 1, the current context window contains a query: "hello how are you?". The most active output node

corresponds to “good”. So “good” will be added to the context window and the process will repeat. Over the next few iterations a trained network might output “good how are you?” A special stop token (not shown in Figure 1) will eventually cause the chat system to stop outputting tokens and wait for the user to make another query.

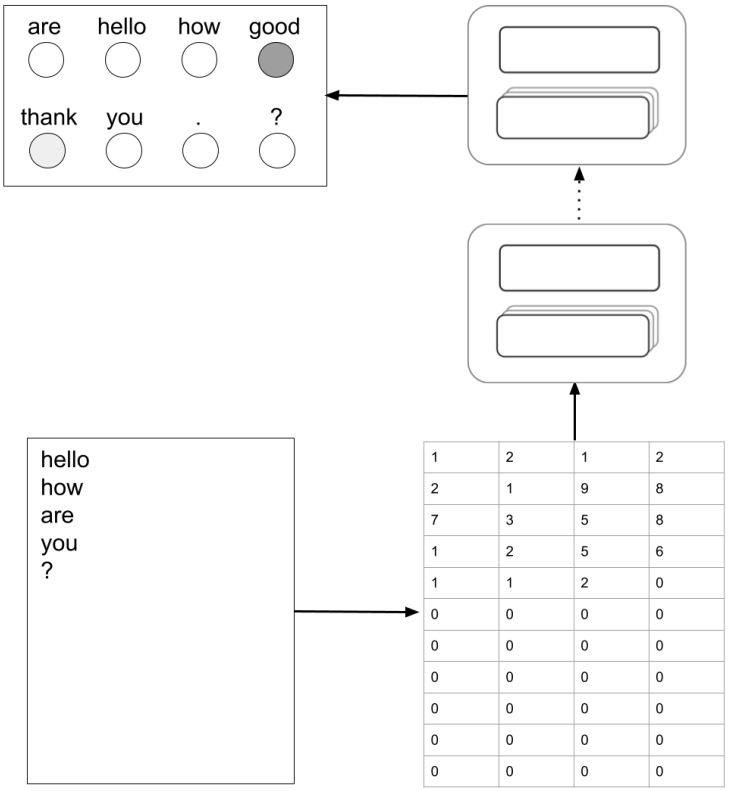

Figure 1: Basic structure of a language model. Tokens in the context window (lower left) are converted to vectors (lower right; notice that unused entries are associated with 0s). These vectors are fed through the network (upper right) and produce an output (upper left) which can be thought of as a probability distribution over possible tokens.

The whole setup is trained using the same backprop algorithm that was popularized in the 1980s. The training data is drawn largely from “common crawl”, which is a massive dataset containing trillions of lines of text scraped from the internet, including published books and

papers as well as material from news sites, social media, blogs, and sundry chatter as well. Inputs correspond to fragments of text, like “The founder of phenomenology was”, and targets correspond to the next token in these fragments, in this case “Husserl”. Over time the network should learn to activate the “Husserl” node more than the “Zebra” or “pillow” node when responding to that sequence of input tokens.

These networks initially struggle to handle novel queries--bits of text they have never seen before--but they suddenly get better after long periods of training (Nanda et al. 2023). This is sometimes called “grokking”. The emergence of grokking is keyed to the development of special circuits inside of LLMs that are detected using the probes and methods of mechinterp research, which in turn suggests LLMs are doing more than simply parroting their training data but are rather developing circuits that allow for generalized responses to novel, unseen inputs.

Figure 2 illustrates some basic features of the internal architecture of LLMs. After tokens are converted into numbers, all information is processed through special layers called “blocks” which contain a collection of “attention heads” and a small feed-forward network or “MLP” (for “multi-layer perceptron”). Frontier networks often have over 100 blocks, and each of these blocks often contains more than 100 heads. All blocks and MLPs read from and write to a common processing bus called the “residual stream”. Roughly speaking each attention head grabs specific information from the residual stream and passes it through the MLP where that information is enriched with stored information, so that as the representation of a token proceeds up through the stream it is progressively equipped with more and more information, which ultimately supports a good guess about the next token.

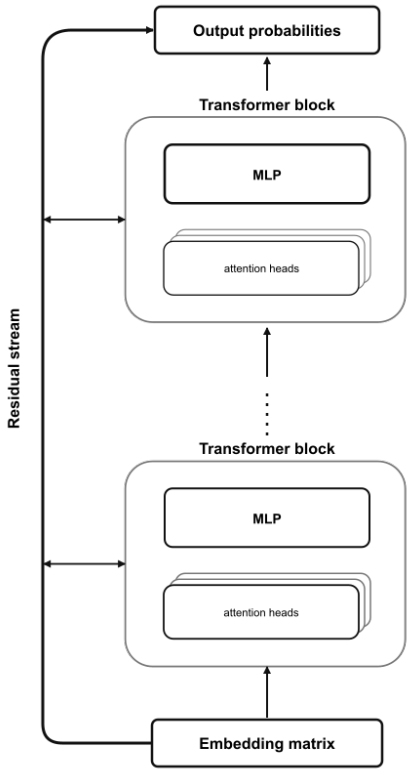

Figure 2: Internal structure of a standard LLM and its transformer blocks. Tokens are converted to numbers by an embedding matrix and transformed by blocks which contain attention heads which filter information and MLPs which enrich that information with factual detail. All information is written to or read from a residual stream.

When these models are trained, they initially speak a chaotic form of internet-like language, what I have elsewhere called “internetese” (Yoshimi, et al., 2025). Conversations wander around and can change in tone radically from moment to moment, shifting from dry factual reporting to abusive online heckler. They develop their distinctive butler-like persona in “fine tuning”, where separate training techniques are used to shape their ability to answer queries in a helpful, pleasing way (though they can in principle be fine-tuned to have other personalities). The behavior of an LLM can be further shaped just by prompting, that is, by adding information to the context window. In fact, whenever you interact with a chat system, a

hidden “system prompt” is added to the context window behind the scenes, which gives the LLM some ground rules for how they should behave (additional hidden changes to the context window occur via dynamic “prompt injections”).

LLMs are usually embedded in a larger ecosystem of tools, what is sometimes called the “harness”, so that in talking to an LLM much more than what I have surveyed above is happening. Separate systems manage security, memory for past chats, and tool calls like internet searches. In some cases, the harness decomposes a task into multiple parts handled by multiple LLM “agents”. For example, when LLMs write code it is not uncommon to have one agent write the code and another set of agents perform an “adversarial review”, where they critique the first LLM’s code and suggest revisions.

4. Basic comparison

Having seen the basics of how LLMs work, we now turn to the main work of this paper, comparing and contrasting LLMs with humans until we isolate a specific human-relevant feature of their circuitry. We start with the “compare” part of the process, emphasizing ways that LLMs are like us.

The first, most obvious comparison to emphasize is that LLMs produce fluent human language. In fact, they pass what has come to be known as a three-party Turing Test, in which a human judge chats with both a human and a machine and tries to guess which is the human (Jones and Bergen, 2026). Turing himself predicted machines would eventually fool the judge

30% of the time, and it is natural to take this to be his threshold for passing (Turing, 1950).[8] Earlier systems routinely and obviously failed three-person Turing tests. However, with the advent of generative AI, chat systems that are prompted to behave like humans (which reduces the giveaway "polite butler" persona) pass. In fact, in a 5-minute test, the best models beat the humans 73% of the time, far ahead of Turing's 30% (Jones and Bergen, 2026). That is, the machine is judged to be human more often than the human is. Whatever your thoughts are about AI, it's hard to deny that this constitutes a major scientific advance. The kinds of linguistic abilities these systems exhibit is mind-bogglingly good in comparison to earlier models.

A related "win" for LLMs is that they surmount a key component of Hubert Dreyfus' decades-long challenge to the possibility of AI, best known from his books *What Computers Can't Do* (1972), and later, *What Computers Still Can't Do* (1992). Dreyfus presents many arguments but a central idea is that traditional computers programmed using symbolic AI couldn't possibly behave in the same kind of open, situationally aware way a human could, because that would require encoding every possible fact relevant to every possible situation into a vast database. The world presents a kind of open texture of situations and humans generally know how to deal with these situations, even when they are quite unusual. One way he makes the point is in terms of Schank and Abelson's story-understanding model (Schank and Abelson, 1977). The model could answer simple questions about a restaurant scene, for example "was any food ordered?" But, Dreyfus notes, however much the programmer has anticipated with their

[8] Turing never actually specifies a threshold, but he did think machines would be able to think and predicted that "in about fifty years' time it will be possible to programme computers... to make them play the imitation game [which can be thought of as a special case of a Turing Test] so well that an average interrogator will not have more than 70 per cent chance of making the right identification after five minutes of questioning" (Turing 1950).

programming, it's always possible to come up with new questions that a human could easily answer but that no programmer has anticipated yet:

> Did the server walk forward or backward?
>
> Did the customer eat his food with his mouth or his ear?

Of course, researchers could add these items to the database, but then Dreyfus could come up with newer, weirder questions. This was sometimes called the problem of commonsense background knowledge. For years Dreyfus had running debates with researchers in the field, in which he would claim AI systems were limited, researchers would claim they could handle the problem, and then Dreyfus would show new limitations in whatever they built (Dreyfus 1992, 2007). I am familiar with the process on a personal level. For over 20 years I have taught philosophy of cognitive science annually, and as part of the class we collectively interrogate the current-best AI chat systems. Until 2017 the best systems would routinely fail to deal with strange questions probing background knowledge, in accordance with Dreyfus' arguments. With the advent of generative AI, the situation has changed. LLMs can handle these strange questions as easily as a human. No doubt if Dreyfus were here, he'd find other problems[9], but to a first approximation, the problem of commonsense background knowledge has been solved.

So the behavioral comparison between humans and LLMs is broadly favorable. There are structural parallels between humans and LLMs as well. At any time, humans have an immediate awareness of what is currently being said, as well as a contextual awareness of what has been

[9] Melanie Mitchell is carrying the Dreyfusian torch (Mitchell and Krakauer 2023), finding the kinds of limitations he emphasized in LLMs, though many of her challenges have been met to date and the status of her critique is rapidly changing (Yoshimi et al. 2025).

said recently, via what Husserl (1991) calls "internal time-consciousness" and what James (1981) called "the specious present". We also have an ability to react appropriately to this temporal context based on a vast reservoir of unconscious knowledge.[10] In a similar way, a network with its billions of weights responds to a context, in a way that prioritizes what has been said most recently (Liu et al. 2024). The current input segment is a bit like what a person is now hearing or saying, the rest of the context window is a bit like a person's temporal awareness of what was said in the recent past, and the network itself is like a person's unconscious background knowledge. So, at a coarse-grained level, the structure of an LLM parallels the structure of human consciousness of language in relation to the unconscious. The idea is illustrated in Figure 3.

[10] Quite a bit of detail is being omitted here. First, the analog in psychology of time consciousness is working memory for a primal impression and for other recent retentions something like what has been called "activated long term memory" (Cowan 2017). Second, while I have linked time consciousness to the *context window* in an LLM, we could also link time consciousness to *early layers of processing* in an LLM, which represent inputs from the context window using neural activations. I believe my core argument applies independently of where this line is drawn, so long as *something* in the LLM's network is analogized to fading retentions in time consciousness as contrasted with fully unconscious background knowledge.

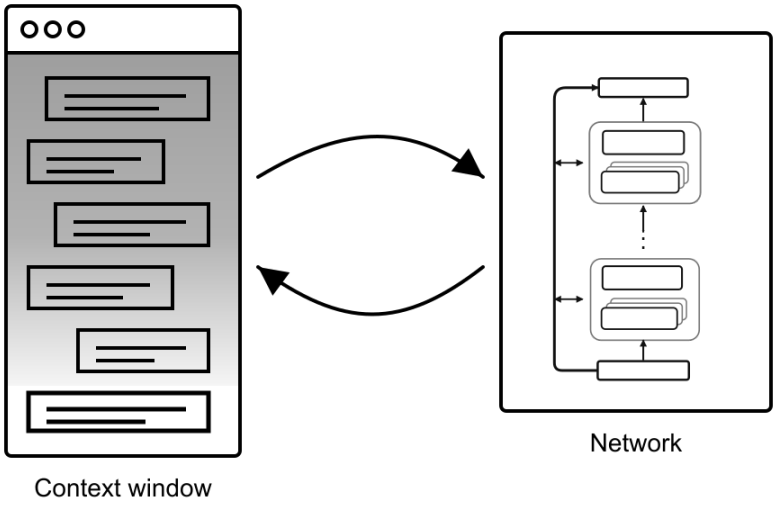

Figure 3: The current input segment in a context window (the query at the bottom of the left panel above) is roughly parallel to what a person is now saying or hearing. The rest of the context window is roughly parallel to a person's temporal awareness of what has recently been said. The network (right panel above) is roughly parallel to a person's unconscious background knowledge.

We now turn to differences between LLMs and humans. Many of these are obvious. LLMs have no agency, no guilt, no emotions, no drives, and no instincts. They don't have mental breakdowns or experience the kinds of internal conflicts humans do. Their memories don't decay over time. They have no individualized sense of themselves and their history, no "narrative" about themselves, in part because they are not embodied. Thus they have no episodic memories in the sense of long-term memories tied to specific life events. Insofar as they have a model of the world it is derived from a kind of statistical average of what's on the internet (Yildirim and Paul 2024). They also know far more than any actual human knows and feel almost perfect in their general knowledge of the world. Compare calculators: they are not good models of human

mathematical reasoning because they never make mistakes, while humans routinely make specific kinds of mathematical mistakes.[11]

We can easily make the LLM pretend to have some of these features. LLMs can act like they have episodic memories or conflicts, but that is not the same as actually having episodic memories or conflicts. To say otherwise is to commit what I will call this the *fallacy of theatrical reconstruction*. Precisely specifying this fallacy is surprisingly difficult (work for another paper!), but I have something like the following in mind: the way a model behaves when it directly reflects its training data, independently of special prompting or fine-tuning, is its "natural state".[12] This is its core architecture and the behaviors of that architecture are its natural behaviors. Anything we have asked it to do just to be more human-like is theatrical reconstruction.

In terms of architecture, there are also striking differences. LLMs are purely feed-forward neural networks, but the brain is massively recurrent. LLMs are disembodied. They sit in a room and read more text than any human could in a single lifetime. Human working memory has a fixed and relatively small size. LLMs have context windows that grow as chat unfolds and becomes extremely large, encompassing millions of tokens or more (an entire 100 page book can be fit in a modern LLM's context). The LLM can "look" at all of its context at each iteration.

---

[11] LLMs are not perfect; they are well-known for their confabulations and hallucinations. But in terms of basic factual recall they are at least analogous to a calculator in their performance relative to humans.

[12] This is clearly an issue that involves prototypical relationships between graded categories and a bit of "I know it when I see it." Both AI and humans are trained and prompted to behave in certain ways in certain contexts. When a child is told to be polite at a social event (the human analog of prompting), we don't typically call this a theatrical production, even if it is not their normal mode of behavior. Similarly for a person trained to be a waiter (the human analog of fine tuning) who behaves much differently waiting tables than they do otherwise. But when someone is told to pretend to have memories that are not theirs, or to fool someone into thinking they are South African by spending a few weeks practicing a South African accent, that is clearly theatrical.

They are a bit like a person who could be shown a 100 page book and immediately "see" all of it in their mind and thus be able to recite any part of it on demand.

Even with all these differences, LLMs are excellent models of language. Just as trained vision systems in industry ended up outperforming custom-built models of the visual system (section 2), so too with language. When early transformer models like "BERT" came on the scene in the late 2010s, they outperformed decades of painstakingly built language systems. Though BERT was not designed to be a scientific model, computational linguists were immediately interested, and within a few years BERT had dominated the field. Some referred to this as BERTology (Rogers, Kovaleva, and Rumshisky 2020). Recall Zipser's dictum: an unrealistic neural network, once trained, behaves a lot like a real cognitive system. The right task demands applied to an abstract circuit produce realistic behaviors.

In fact, across a broad range of domains, LLM performance is comparable to human performance. Experimental studies of LLMs have found human-like reading times, semantic valence, syntax, grammaticality, pronoun binding over long distances, and semantic spaces (Chang and Bergen 2024; Schrimpf et al. 2021; Yoshimi et al. 2025). In some cases these studies take old data collected for humans, such as ratings of words on various qualities (e.g. how "concrete" or "familiar" different words are), then feed that same data to LLMs, and find strong correlations between the two (Kello, Bruna, and Thao 2026). Moreover, the internal activations of LLMs when they read a text are similar to neural activity in the human brain when it reads the same text. Show a human a text and measure their brain and measure an LLM as it reads the same text, and some specific layers in the LLM are highly predictive of brain activity in language-processing parts of the brain (Caucheteux and King 2022).

Of course, there are many details and controversies surrounding these studies, but at a high-level it seems clear that LLMs capture *something* about what humans do when reading, hearing, and speaking.

5. The comparison of interest

We have seen that LLMs, despite their obvious differences from humans, are still our best current models of language processing and are predictive of neural activity. That suggests that they might somehow be relevant to the phenomenology of language and unconscious access. In this section I make the case that they are.

I will focus on a pair of basic features of LLMs, one concerning ability, the other concerning format. Together they describe a "meaning-saturated appropriate access circuit". This circuit was implicit in the discussion above but is being isolated and separately described here.

Consider first your ability to recall relevant information when speaking or thinking. I will call this "appropriate access". When we speak with others, we pull out just the right things from unconscious background knowledge and say them at just the right time. We speak fluently and naturally. As we have seen, LLMs can do this as well. They pass the Turing Test and solve the problem of background knowledge. They do this by "contextually assembling", out of their internal networks, just the right information needed to plausibly continue any conversation (Kello, Bruna, and Thao 2026). So this much of the phenomenology of language and unconscious access can be explained by the circuitry of LLMs.

Second, LLMs have the right format to do the job. From phenomenological considerations alone, we can make certain inferences about the structure of unconscious

knowledge. It is some kind of vast and non-propositional source of abilities, a collection of "horizons" and "sedimented knowledge" that determines how we react to situations (Husserl 1973, 2001). The details of its structure are hard to delineate precisely, but we can with some confidence say that it is *not* a collection of unconscious propositions. Language is too brittle to describe every fact that might be relevant to a situation. This is basically Dreyfus' argument, and it is supported by the fact that symbolic AI never achieved human-level performance in natural language tasks.

LLMs have an appropriate format for implementing commonsense background knowledge. They are non-propositional: they were not programmed, but were grown through training. We cannot read meanings off of them the way we can read items in a database. The meanings within them overlap in complicated ways, suggestive of the "open textures" and "horizons" phenomenologists describe. In fact, thanks to the principle of "linear superposition", a set of nodes can support far more meaningful structure than their number would suggest (Beckmann and Queloz 2026). This is because meanings correspond to *directions* in a space, which in turn correspond to overlapping patterns of node activation rather than to individual nodes. In fact, a set of 100 neurons can support tens of thousands of accessible feature representations (Garg, Kleinberg, and Peng 2026). Thus, neural networks can be *super*-saturated with meanings, containing far more entangled features than the number of neurons within them. These directions can be instantly accessed and assembled thanks to the way networks learn to channel and filter information.

In fact, LLMs are a lot like the holographic memory systems Dreyfus once championed. As Dreyfus said in an interview, "There are already holographic models of pattern recognition where you would have, say, fifty thousand typical driving situations stored. When the current

input was like one of them, it would just pull that one out. It wouldn't have to scan all the others. You've got to have a model like that, because it turns out that these intuitive, skillful responses are practically instantaneous" (Dreyfus, quoted in The Intuition Network n.d.).

So, LLMs produce behaviors that are comparable to humans when they use language. They assemble the right information in the right way, nearly instantly, without relying on propositional databases. That's the important "not nothing" I have tried to isolate in this paper. In the conclusion, I will argue that this circuit could be the basis of important further studies, which highlight other ways LLMs shed light on neuro-phenomenology.

Before moving on, there are a few more specific differences to be considered within the domain of unconscious access and language. In a chatbot, "appropriate access" circuits are used in a very non-human way. After all, they were designed to be helpful assistants, not realistic models of humans. We can consider differences in behavior, architecture, and development.

The dynamics of language in its relation to the unconscious is in many ways different between humans and LLMs. In conversation, humans stutter and halt. We have trouble finding the right word. We get pulled into sidelines and are led astray by side thoughts. We pause as we decide how to frame a point. We vary in how much we understand things and are more fluent when we speak on familiar topics. Certain kinds of pressure make us nervous and change how we speak. Sometimes ideas hit us days later, after having slept. By contrast, LLMs always behave in a uniform, perfect kind of way.[13] They speak fluently and reliably on demand, at a consistent speed, and in pretty much all circumstances. They can be tuned or prompted to make mistakes or show gaps in their understanding, but to take such behaviors as evidence of these

[13] LLMs can vary in their reaction time, for example in proportion to how many failed solutions are attempted by the LLM before a working solution is found, but this is more analogous to an extended reasoning process than to the milliseconds-level process of halting and stuttering through a conversation.

qualities is to commit the fallacy of theatrical reconstruction. Telling a “perfect” network to stutter or halt or act conflicted is not the same thing as actually stuttering or halting or being conflicted.

These differences are rooted in fundamental architectural differences. Consciousness is often thought to be structured into a focus of attention and a fringe or periphery of inattention, a “field of relatively unarticulated, vague experience” (Mangan 1993). When we talk or think we generally have a focal sense of what is being said but also a vague sense, in the fringe, of different directions we might take in a conversation. Bruce Mangan has, perhaps more than anyone else, considered these ideas from a perspective that is sensitive both to phenomenology and to information processing. On his account

> ...fringe experiences (1) work to radically condense context information in consciousness; (2) are vague because a more explicit representation of context information would overwhelm consciousness' limited articulation (processing) capacity; (3) help mediate retrieval functions in consciousness; and (4) contain a subset of monitoring and control experiences that cannot be elaborated in focal attention and are "ineffable" (Mangan 1993, 89).

A specific variety of fringe experience Mangan emphasizes is what he calls “rightness”, which “functions as a summary index of cognitive integration, representing, in the fringe, the degree of positive fit between a given conscious content and its parallel, unconsciously encoded context.” Rightness supports the dynamics described above. When having a conversation or even just thinking about something, we generally have a vague sense of ways to continue the conversation or thought process. When a single item feels most right, we simply follow that direction and the conversation or thought process feels smooth and natural. In some cases, we are almost pulled through a chain of associations. The operation of an LLM is probably closest to cases like this,

where a human confidently discourses on a familiar topic and from moment to moment the brain delivers up just the right thing to say.

But in other cases we are less sure about what to say, and multiple options might feel equally right. In those cases we pause and ponder and sometimes simply give up and move on to another topic. There is nothing like this in an LLM.[14] There is no dynamic of considering a few options, having that reflected in response times, and then choosing the one that seems best: there is just the steady, uniform march of information through the layers of the network as tokens are processed. There are no explicit coherence computations that I am aware of in an LLM. There are shadows or echoes of these things[15], but nothing like a fringe is an explicit part of its architecture.

A final difference concerns the development of background knowledge. As noted above, human knowledge is organized around our embodied experience of the world, which unfolds over the days and years of our lives. Everything we know, everything in our unconscious

---

[14] While finalizing this article Anthropic published a report (Anthropic 2026) claiming to find a key architecture associated with consciousness, the "global workspace" (Baars 198 8), in the middle layers of LLMs. These layers "maintain a small, privileged set of representations that they can report, manipulate, and reason with, amidst a much larger volume of processing that they cannot". This is suggestive of a circuit that approximates some features of information processing relating to consciousness, but the authors rightly emphasize the many ways that LLMs *don't* implement this architecture: "Despite these similarities, we do not claim that language models reproduce the full architecture global workspace theory ascribes to the brain—specialized, encapsulated processors competing for entry to a workspace that broadcasts back to them through recurrent connections... Several of those features have no clean analog in a transformer-based language model: there are no obviously separable input processors, and the broadcast we document occurs within a single feedforward pass rather than through recurrent loops. Moreover, although we observe some degree of competition for access... it is unclear whether this mirrors the sharp, competitive 'ignition' that characterizes workspace entry in the brain."

[15] Here are some mechanisms that are analogous to fringe structure insofar as they involve selective filtering of information, and perhaps on closer inspection might involve some analog of a rightness computation. The filtering mechanism built into attention can be thought of scoring information in the residual stream for how relevant it is to the current processing at a layer. The mixture of expert systems used in some LLMs learn to route information to specialized subnetworks may route information based on something like rightness(Fedus, Zoph, and Shazeer 2022).

background knowledge, is tied to these personal experiences. All of a person's knowledge is wound up in the grand narrative of their life.

Husserl has a long and fascinating discussion of these processes in *Husserliana* 42 (2014), a set of unpublished notes Husserl took concerning the phenomenological status of sleep, death, and other structures that we are aware of when conscious but that we don't actually live through. He develops the idea that current experiences first sink into the retentions of time consciousness, where they recede and are finally incorporated or "sedimented" into the "darkness" of the unconscious. As Husserl says

> The process of retention is a process of darkening …of continuous de-strengthening, whose limit is unwakefulness, forcelessness (Husserl 2014, Text 3, p. 36)
>
> The formal determination "sedimented" refers in each case to what lies below that zero. (Husserl 2014, Text 3, p. 40)

Husserl admits that the way these fading memories intermingle when they are sedimented into the unconscious is unclear:

> Finally we arrive at the limit of zero… what then becomes of all the other zero residues that, unified through their origin, are somehow connected... within the sedimented unconscious… What makes the metaphor of sedimentation useful? What makes the supposed layering upon one another akin to a fusion within the general night? (Husserl 2014, Text 3, p. 63).
>
> What, then [happens] in the darkness of sedimentation? (Husserl 2014, Text 3, p. 63)

Though Husserl's analysis leaves many questions unanswered, what is clear is that LLMs have none of this. LLMs do not develop their knowledge as mobile agents in the *umwelt*. They

are force-fed the entirety of written knowledge while sitting in massive data centers. They have no narrative self-understanding. They have no specific experiences of events that happened to them at specific times. LLMs can invent autobiographical details and even pass the Turing Test but taking that as evidence of a human-like unconscious is to commit the fallacy of theatrical reconstruction.

6. Conclusion

LLMs use a non-propositional super-saturated meaning structure that supports the phenomenology of appropriate access. These circuits are in many ways unrealistic: they lack emotions or drives; they are developed in a non-human way; they know just about everything; and they maintain a constant, uniform demeanor. But the circuits themselves are in some ways realistic as models of unconscious access, and could explain some aspects of the phenomenology of language

These insights could be the basis of further studies of the phenomenological significance of LLMs. The success of AI in doing even the limited things I have focused on goes beyond what happens in a single layer. Something about the multi-layer arrangement of these circuits and the combination of attention heads for filtering information and MLPs for storing information is suggestive of the organization of unconscious knowledge. Due to space considerations, I have omitted detailed considerations of these features, but will briefly consider them here. In terms of layered organization, deep networks for vision offer an instructive comparison. We see things as having edges and textures and contours and other ineffable qualities such as the “snoutiness” of dog noses, and this phenomenology is both explained and

informed by the circuitry of deep vision networks (Yoshimi forthcoming .). Similarly here. Structures responsive to part of speech, numbers, sentiment, dates, geography, and many other concepts have been found in LLMs (see section 2). These structures are organized in a layered way. Earlier layers focus on spelling and punctuation; middle layers on semantics and reference resolution; later layers on factual recall and reasoning. Some parts of the network (the "MLPs") specialize in storing information and others (the attention heads) in retrieving and shuttling information around and deciding which information a given layer should focus on. All of this is tantalizingly suggestive of human cognition and phenomenology. It does seem that when we have a conversation that a multitude of meaningful features are present but interwoven: our sense of the syntax and the prosody of the language, of whether a conversation is humorous or serious, of which facts matter to our current theme, of how detailed we should be, of whether we have been understood, and so forth.

It is also worth noting that many variations on the basic LLM architecture described in section 2 are being tried, for example, more-efficient models that can run on a personal computer by using more localized computations, models that more actively gate information, and models that maintain small state space representations to model dynamic processes (Gu and Dao 2024; Fedus, Zoph, and Shazeer 2022; Amini et al. 2025). This area is extremely active. Some of these variations could be more relevant to neurophenomenology than the massive LLMs that are currently industry standard.

But ultimately none of this will be enough. For these circuits to be organized into more adequate cognitive models, I believe they would have to be wired up and configured in a completely different way. What do I think a more realistic use of these circuits look like? Here is an admittedly vague and programmatic start. We can imagine a model with a relatively small,

fixed-size context window. The window would have to have some explicit distinction between what is immediately available in the focus of attention and what has faded into something like activated long term memory. Such a distinction could recover some of the structure of human time-consciousness. The model would have to be embodied in a real or virtual environment and would have to have a humanoid body. The AI would have to be called by name and have real experiences with others (Turing imagined just this situation, and worried that the children-AI would be teased). Work along these lines has begun (Driess et al. 2023). The network itself could be structured into a main agent and a confederation of unconscious agents. The main agent would receive signals from the unconscious agents and would use a "rightness signal" to select information for further processing by the main agent.[16] To be clear, I don't think such a model would actually be conscious[17], but such a model could capture more of the *structure* of human phenomenology than existing AI does.[18]

Bibliography

Amini, A., Banaszak, A., Benoit, H., Böök, A., Dakhran, T., Duong, S., Eng, A., Fernandes, F., Härkönen, M., Harrington, A., Hasani, R., Karwa, S., Khrustalev, Y., Labonne, M., Lechner, M., Lechner, V., Lee, S., Li, Z., Loo, N., Marks, J., … Rus, D. (2025). *LFM2 technical report* arXiv. https://doi.org/10.48550/arXiv.2511.23404

Anthropic. (2026, July 6). *A global workspace in language models*. https://www.anthropic.com/research/global-workspace

Baars, B. J. (1988). *A cognitive theory of consciousness*. Cambridge University Press.

[16] A collection of LLMs wired up along these lines would still be a far cry from the neural networks of an actual brain, and I believe those differences matter to the structure of human phenomenology. I am just indicating the direction in which I believe a more adequate theory lies.

[17] I incline towards substrate-focused or "biological naturalist" theories as against what are now called "computational functionalist" theories (Seth 2025), though none of my arguments here depend on that inclination.

[18] Helpful feedback was provided by Bruce Mangan, Tyler Marghetis, Michael Spivey, and an audience at the Southwest seminar in Continental Philosophy.

Beckmann, P., & Queloz, M. (2026). Mechanistic indicators of understanding in large language models. *Philosophical Studies*. Advance online publication. https://doi.org/10.1007/s11098-026-02513-1

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? 🦜 In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Burrell, J. (2016). How the machine "thinks": Understanding opacity in machine learning algorithms. *Big Data & Society, 3*(1), 1–12. https://doi.org/10.1177/2053951715622512

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology, 5*, Article 134. https://doi.org/10.1038/s42003-022-03036-1

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics, 50*(1), 293–350. https://doi.org/10.1162/coli_a_00492

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Cowan, N. (2017). The many faces of working memory and short-term storage. *Psychonomic Bulletin & Review, 24*(4), 1158–1170. https://doi.org/10.3758/s13423-016-1191-6

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience, 24*(7), 431–450. https://doi.org/10.1038/s41583-023-00705-w

Dreyfus, H. L. (1972). *What computers can't do: A critique of artificial reason*. Harper & Row.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. MIT Press.

Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence, 171*(18), 1137–1160. https://doi.org/10.1016/j.artint.2007.10.012

Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., & Florence, P. (2023). *PaLM-E: An embodied multimodal language model* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2303.03378

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., . . . Olah, C. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*. https://transformer-circuits.pub/2021/framework/index.html

Fazi, M. B. (2024). The computational search for unity: Synthesis in generative AI. Journal of Continental Philosophy, 5(1), 31–56. https://doi.org/10.5840/jcp202411652

Fedus, W., Zoph, B., & Shazeer, N. (2022). Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research, 23*(120), 1–39. https://www.jmlr.org/papers/v23/21-0998.html

Freud, S. (1966). Project for a scientific psychology (J. Strachey, Trans.). In J. Strachey & A. Freud (Eds.), *The standard edition of the complete psychological works of Sigmund Freud: Vol. 1. Pre-psycho-analytic publications and unpublished drafts* (pp. 281, 295–387). Hogarth Press. (Original work written 1895)

Gallagher, S. (2012). On the possibility of naturalizing phenomenology. In D. Zahavi (Ed.), *The Oxford handbook of contemporary phenomenology* (pp. 70–93). Oxford University Press.

Garg, N., Kleinberg, J., & Peng, K. (2026). How many features can a language model store under the linear representation hypothesis? In S. Hanneke & T. Lattimore (Eds.), *Proceedings of the 39th Conference on Learning Theory* (Vol. 336, pp. 5358–5376). Proceedings of Machine Learning Research. https://proceedings.mlr.press/v336/garg26a.html

Gu, A., & Dao, T. (2024). *Mamba: Linear-time sequence modeling with selective state spaces* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2312.00752

Gurnee, W., & Tegmark, M. (2024). Language models represent space and time. In *Twelfth International Conference on Learning Representations*. https://arxiv.org/abs/2310.02207

Gurwitsch, A. (1964). *The field of consciousness*. Duquesne University Press.

Humphreys, P. (2009). The philosophical novelty of computer simulation methods. *Synthese, 169*(3), 615–626. https://doi.org/10.1007/s11229-008-9435-2

Husserl, E. (1973). *Experience and judgment: Investigations in a genealogy of logic* (J. S. Churchill & K. Ameriks, Trans.; L. Landgrebe, Ed.). Northwestern University Press. (Original work published 1939)

Husserl, E. (1991). *On the phenomenology of the consciousness of internal time (1893–1917)* (J. B. Brough, Trans.). Kluwer Academic Publishers.

Husserl, E. (2001). *Analyses concerning passive and active synthesis: Lectures on transcendental logic* (A. J. Steinbock, Trans.). Kluwer Academic Publishers. https://doi.org/10.1007/978-94-010-0846-4

Husserl, E. (2014). *Grenzprobleme der Phänomenologie: Analysen des Unbewusstseins und der Instinkte. Metaphysik. Späte Ethik (Texte aus dem Nachlass 1908–1937)* (R. Sowa & T. Vongehr, Eds.). Springer. https://doi.org/10.1007/978-94-007-6801-7

The Intuition Network. (n.d.). *Mind over machine with Hubert Dreyfus, Ph.D.* [Interview transcript]. https://www.intuitionnetwork.org/txt/dreyfus.htm

James, W. (1890). *The principles of psychology* (Vols. 1–2). Henry Holt and Company.

James, W. (1981). *The principles of psychology* (F. H. Burkhardt, general editor; F. Bowers, textual editor; I. K. Skrupskelis, associate editor; Vols. 1–3). Harvard University Press.

Jones, C. R., & Bergen, B. K. (2026). Large language models pass a standard three-party Turing test. *Proceedings of the National Academy of Sciences*, 123(21), e2524472123. https://doi.org/10.1073/pnas.2524472123

Kello, C. T., Bruna, P., & Thao, K. (2026). Contextual assembly of lexical functions in large language models. *Behavior Research Methods, 58*(1), Article 19. https://doi.org/10.3758/s13428-025-02898-7

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2024). Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics, 12*, 157–173. https://doi.org/10.1162/tacl_a_00638

Mangan, B. (1993). Taking phenomenology seriously: The "fringe" and its implications for cognitive research. *Consciousness and Cognition, 2*(2), 89–108. https://doi.org/10.1006/ccog.1993.1008

Mitchell, M., & Krakauer, D. C. (2023). The debate over understanding in AI's large language models. *Proceedings of the National Academy of Sciences, 120*(13), Article e2215907120. https://doi.org/10.1073/pnas.2215907120

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=9XFSbDPmdW

Niu, J., Lu, W., & Penn, G. (2022). Does BERT rediscover a classical NLP pipeline? In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 3143–3153). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.278/

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020). Zoom in: An introduction to circuits. *Distill*. https://doi.org/10.23915/distill.00024.001

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*, 842–866. https://doi.org/10.1162/tacl_a_00349

Rumelhart, D. E., & McClelland, J. L. (Eds.). (1986). *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1. Foundations*. MIT Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Lawrence Erlbaum Associates.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J. B., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences, 118*(45), Article e2105646118. https://doi.org/10.1073/pnas.2105646118

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems, 1*, 145–168.

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*. Advance online publication. https://doi.org/10.1017/S0140525X25000032

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., & Henighan, T. (2024). Scaling monosemanticity: Extracting interpretable features from Claude 3

Sonnet. *Transformer Circuits Thread*. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1452

Tigges, C., Hollinsworth, O. J., Geiger, A., & Nanda, N. (2024). Language models linearly represent sentiment. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP* (pp. 58–87). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.blackboxnlp-1.5

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Varela, F. J. (1996). Neurophenomenology: A methodological remedy for the hard problem. *Journal of Consciousness Studies, 3*(4), 330–349.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*, 5998–6008.

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences, 111(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yoshimi, J. (2016). Prospects for a naturalized phenomenology. In D. O. Dahlstrom, A. Elpidorou, & W. Hopp (Eds.), *Philosophy of mind and phenomenology: Conceptual and empirical approaches* (pp. 287–309). Routledge.

Yoshimi, J. (2022). Phenomenological psychology as philosophy of mind. In H. Jacobs (Ed.), *The Husserlian mind* (pp. 446–458). Routledge. https://doi.org/10.4324/9780429243790-43

Yoshimi, J., Tosi, Z., Hotton, S., Udell, D., Beckmann, P., Cain, E., Gordon, C., Noelle, D. C., Bruna, P., & Meyer, T. (2025). *Neural networks in cognitive science* (Version 2025.3) [Open textbook]. https://downloads.jeffyoshimi.net/NeuralNetworksCogsci.pdf

Yoshimi, J. (forthcoming). Computational phenomenology. In H. A. Wiltsche (Ed.), *The Oxford handbook of phenomenology of science*. Oxford University Press.

Zipser, D. (1992). Identification models of the nervous system. *Neuroscience, 47*(4), 853–862. https://doi.org/10.1016/0306-4522(92)90035-Z

Zhou, T., Fu, D., Sharan, V., & Jia, R. (2024). Pre-trained large language models use Fourier features to compute addition. *Advances in Neural Information Processing Systems, 37*. https://doi.org/10.52202/079017-0792