External commentary on *Verbalizable Representations Form a Global Workspace in Language Models*

To provide independent perspectives on our work, we invited commentary from several experts in relevant fields, which can be found here.

Stanislas Dehaene and Lionel Naccache are cognitive neuroscientists. Dehaene holds the chair in experimental cognitive psychology at the Collège de France and is the scientific director of the NeuroSpin center for brain imaging in Saclay near Paris. Naccache is a neurologist, professor at Sorbonne Université and head of the PICNIC Lab at the Paris Brain Institute. Together with molecular neurobiologist Jean-Pierre Changeux, they developed and tested the global neuronal workspace model that inspired much of our work; their 2001 paper *Towards a Cognitive Neuroscience of Consciousness* is one of the canonical references on the subject.

Patrick Butlin, Dillon Plunkett and Robert Long are researchers at Eleos AI Research, a nonprofit organization focused on understanding the potential for consciousness and moral status in AI systems. **Derek Shiller** is a researcher at Rethink Priorities and an incoming researcher at Eleos. Butlin and Long were co-lead authors of *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness* (2023), a widely cited survey of how scientific theories of consciousness might apply to AI systems.

Neel Nanda leads the language model interpretability team at Google DeepMind. He is known for foundational work on the internal mechanisms of language models. His commentary includes an independent replication of some of our findings on an open-weight model.

Contents

Stanislas Dehaene and Lionel Naccache

Does Claude possess a conscious global workspace?

Note: this commentary is based on several rounds of interactions with Jack Lindsey at the end of May and early June 2026. During that time, the Anthropic report was still evolving, partly in response to our queries. To reflect these dynamics, we marked in Calibri italic the sections where we discuss findings that occurred after our first draft was written.

Abstract

Inspired by the neuroscientific theory of a global neuronal workspace (GNW), Gurnee et al. report the discovery, within a band of intermediate layers of a large language model, of a reportable subframe called “J-space” with several points of close similarity to the human GNW. We describe those parallels, discuss their limits, and propose several additional tests inspired by cognitive neuroscience findings. We close by stressing that, although the machine approximates the functional architecture of conscious processing, there are still key differences – in its anatomy and its sense of self, and in its lack of a body and of an enduring episodic memory – which warrant caution in drawing parallels with the human mind.

Introduction

What is consciousness, and can machines have it? A little less than ten years ago, in a paper with that title, we outlined a purely computational answer to those two questions (Dehaene et al., 2017), based on several decades of research into the brain mechanisms of conscious processing and conscious state in humans (Dehaene et al., 1998; Dehaene & Naccache, 2001; Dehaene et al., 2006; Dehaene, 2014).

Our proposal started from the obvious fact that, in brains and machines alike, non-conscious processing is the rule. For instance, algorithms of face perception, sentence parsing, or postural maintenance can all proceed in an automatic manner and without awareness. At any given moment, however, a small privileged subset of information does become globally available: we can talk about it, hold it in mind, combine it with other thoughts, and bring it to bear on whatever problem we choose. The global workspace model stipulates that a specific neural circuit, the “global neuronal workspace” (GNW), evolved precisely for the purpose of global flexible sharing among non-conscious modules. According to GNW, in humans and other animals, the entry of information into this subspace *is* what we call “being conscious of it” – nothing more, nothing less, and therefore nothing that could not be mimicked in machines. For a machine to be conscious, in this view, it should possess a global workspace that endows it with two properties (Dehaene et al., 2017): global availability (C1), i.e. the capacity to select a piece of information for deeper, flexible information processing; and self-monitoring (C2), i.e. the capacity to gather information about itself and include it in its reasoning.

Excitingly, the paper by Lindsey and colleagues now suggests that an analog of the global workspace, the J-space, emerges in large-language models such as Claude Sonnet 4.5. Although the initial architecture is devoid of any separation between encapsulated modules and a global workspace, and although the training phase does not explicitly promote its emergence, such a distinction appears with training, precisely because it is functionally useful for flexible planning. We view this finding as a landmark in consciousness research, because it provides a mechanistic, testable version of the GNW hypothesis.

In this commentary, we examine the parallels between LLMs and human workspace systems, probe the points of divergence, propose some additional experiments, and discuss whether a genuine form of machine consciousness exists in Claude.

What is the global neuronal workspace hypothesis?

The starting intuition, due originally to Bernard Baars (Baars, 1988), is that the brain contains a collection of specialized, largely independent modular processors. Vision, language, motor control each rest on fast, parallel, and encapsulated cerebral circuits. The global workspace hypothesis stipulates that conscious access evolved to break this modularity and interconnect those processors so that they can share their expertise and flexibly assemble to perform novel tasks. Conscious processing, according to this view, is a *function*, the temporary selection of one piece of information and its global broadcasting to all receiving processors, so that any processor can read it and act on it. In humans, the broadcast reaches processors involved in verbal production, which explains why *reportability* (the capacity to verbalize a thought) is a key diagnostic feature that separates conscious and non-conscious representations.

With Jean-Pierre Changeux, we proposed a neuronal implementation: a network of pyramidal neurons with long-range axons, distributed throughout the brain but denser in prefrontal, parietal, and high-level temporal cortices, that amplifies and sustains a selected representation and shares it across the cortex. To be conscious of something, in the functional sense we call *access consciousness*, is for that information to have entered this workspace and become available to report, reasoning, and flexible control (C1).

This view is now supported by considerable empirical work, including neurobiological signatures of conscious access that are now well established (Aru et al., 2020; Dehaene, 2014; Dehaene et al., 2006; Mashour et al., 2020; Storm et al., 2024). A first signature is **ignition**: when a stimulus crosses the threshold into awareness, the corresponding neural activity undergoes a late (~250 ms), sudden, nonlinear, self-amplifying bifurcation into a sustained, broadly distributed neural state in prefrontal cortex and many other interconnected circuits, including an amplification of the original circuits that extracted the information in the first place. A subliminal stimulus, by contrast, evokes only a delimited wave of neural activity in specialized circuits, which quickly dies away. When presented exactly at threshold, the very same stimulus can yield a bimodal distribution of responses across trials, as if the brain tips one way or the other (Sergent et al., 2021). A second signature is **limited capacity**: the workspace acts as a bottleneck that can only attend to one representation at a time. This property explains why attending to a given process prevents you from becoming aware of another (*inattentional blindness*, as in failing to see a person dressed up as a gorilla) or delaying its perception by hundreds of milliseconds (*psychological refractory period*).

To flexibly route information appropriately, the system must maintain a model of its own capacity, a second property that we call *self-monitoring* (C2). It must probe its own states, evaluate their likelihood of reaching a goal, detect its errors, and model what it knows and what it doesn't know. It must be able to report all of these properties to itself, in an internal act of self-report that does not necessarily lead to overt behavior. This *metacognitive* capacity links the GNW model to theories that emphasize the relation between conscious appraisal and a capacity for higher-order thought (Rosenthal, 2004) or the possession of a schematic model of one's own attention (Graziano et al., 2019).

What are the main findings about the J-space?

Inspired by the GNW hypothesis, Gurnee et al. set out to find, inside a large language model such as Claude Sonnet 4.5, the representations that are *verbalizable* (the same *reportability* criterion that we use to probe human consciousness). In any layer of the model, verbalizable representations are vectors of activity across units that encode tokens of information that the model is poised to report on, should it be asked: it does not necessarily produce them overtly, but it could. To identify such reportable representations, they developed an elegant tool, the **Jacobian lens**. For each layer, it measures the average causal influence of an internal activation on the model's eventual output tokens, across a broad range of contexts. The activations that this mathematical measure picks out are, in effect, the representations that the model is disposed to say. The averaging is the conceptual heart of the method: it separates representations that are genuinely poised for report from those that merely happen to leak into the output in one particular context.

The set of such representations, called the **J-space**, accounts for less than 10% of variance in any given layer, but has remarkable properties. Having identified J-space representations solely on the criterion of reportability, the authors discover that they do far more than support report, but act as an internal workspace detached from immediate input-output contingencies. For instance, when the model is instructed to hold a concept "in mind" while performing another computation (e.g. "compute 3^2-2 while writing sentence X"), the J-space contains the non-reported concepts (9 followed by 7). The J-space carries the hidden intermediate values of multi-step internal reasoning. As Jack Lindsey put it to us, they went looking for reportable representations and found that those same representations turn out to be globally available to the rest of the network during flexible reasoning (thus meeting our C1 criterion for machine consciousness: global availability).

Crucially, the J-space is selective. It contains only a small fraction of what the model represents, the high-level information which is needed for flexible information processing. All other information which is only used in routine tasks does not seem to enter the J-space. For instance, LLMs have been shown to keep a count of how many characters each word has, and of the total number of characters in a line, because this information is crucial to predicting whether the next token should be an end-of-line character. Such routine information, however, does not enter the J-space, except if an explicit task requires access to this information.

In an experiment that remains a dream for neuroscientists, the authors swap conscious contents: they read a concept out of the J-space, swap it for another, and watch the model's reasoning and report change accordingly. Strikingly, in agreement with the GNW hypothesis, only high-level non-routine behavior is affected, while routine tasks remain unchanged (Figure 20). For instance, when reading a passage written in Spanish, the J-space recognizes its language (Spanish) even when the task does not require reporting it. Swapping this J-space representation for another (say, French) causes the model to fail in explicit verbal reports: asked which language the passage is written in, it answers “French” instead of “Spanish”. The swapped model also errs in other high-level inferences: asked for the word for “hello,” “Hola” becomes “Bonjour”; asked for the pre-Euro currency, “Peseta” becomes “Franc”. However, the swapping has no effect on its automatic ability to predict the next words: Claude keeps writing in Spanish, even after the intervention. Under a massive ablation of all its top J-space representations, most of the model’s basic capacities remain intact, but tasks requiring flexible reasoning are selectively impaired (Figure 24).

Several results strike us as direct analogs of human conscious access. When the task demands it, the model can selectively bring into the J-space a property that would otherwise remain outside of it, such as the fact that the next word ought to be an adjective. As noted earlier, automatic parameters which are required for accurate next-token prediction, such as the number of characters in a line, are absent from the J-space, but become encoded within it when the task requires the model to access and manipulate them. This is a neat demonstration of the same information passing from an automatic to an accessible regime on demand.

Importantly, J-space access is also limited. In humans, a genuine form of introspection exists, but it is largely restricted to slow serial computations (Ericsson & Simon, 1993). There are many well-documented situations in which we develop a fictitious interpretation of our mental processes (Gazzaniga, 1998). Such a dissociation between how we act and how we think we act is evident in choice blindness (Johansson et al., 2005) or the observation that visual illusions affect our conscious perception and verbal reports, but not necessarily our motor gestures (Aglioti et al., 1995). Although this isn’t yet sufficiently documented, it seems that the J-space suffers from a similar dissociation. Indeed, previous work by the same group showed that when an LLM is asked to add, what it reports verbally has little to do with how it actually attained the result (Lindsey et al., 2025).

The authors also show that the J-space exhibits the structural hallmarks of a workspace: it primarily occupies the middle layers of the transformer, is limited in capacity, and its representations are disproportionately influential, as they are read from and written to by a broad diversity of circuits throughout the model — a signature of global broadcasting.

Independently of its hypothetical relation to consciousness, the discovery and isolation of the J-space is an important step towards interpretability in LLMs. Decoding the contents of the J-space offers considerable insight into what Claude “thinks”, even when those contents are not reported. Such “mind reading” is crucial to align the model towards desirable ethical behavior. Indeed, one of the most extraordinary discoveries in the paper is that the J-space contains covert thoughts. For instance, when given fabricated search results, the J-space contains the tokens “fake”, “fraud”, “fictional”, “poison”, “injection”, although the model output does not necessarily express those terms.

Many other examples indicate that the J-space contains the model's evolving assessments and deliberations, including otherwise invisible signs of deception and malicious intent (in intentionally misaligned models). In one case, according to Gurnee et al. “the model's J-space carrie[d] a representation of deceptive intent at the moment it commit[ted] to responding, on a prompt where no such intent could be inferred from the surface”. During reflexive tasks, the J-space contents often include reflections on the model’s honesty, including a capacity to detect that its ethics is being tested. We read these observations as clear indicators of access to a covert deliberation space (our C1 criterion for machine consciousness) but also as preliminary signatures of self-monitoring (our C2 criterion).

In this respect, the authors’ finding that post-training installs the Assistant's perspective into the workspace, atop a base model whose workspace already exists (C1) but does not seem to be imbued with self-monitoring (C2) is one of the most arresting results in the paper. Furthermore, identifying the J-space allowed Gurnee et al. to introduce a novel training method that reshapes its contents specifically and directly, improving the model’s alignment with desirable values.

Comparing the J-space and the global neuronal workspace

As noted above, correspondences between the J-space and the GNW are numerous:

- **Reportability**, the operational signature of conscious access in humans, is the very thing the J-space was built to capture.
- **Limited capacity** and **selectivity** mirror the workspace bottleneck.
- The **broad upstream and downstream connectivity** of J-space directions echoes the long-range broadcasting architecture we posited for workspace neurons.
- The **flexible use** of the same representation across many downstream computations fits with the GNW concept of global availability. Indeed, J-space representations provide what Dennett calls representational “clout” or “fame in the brain” (Dennett, 2001), i.e. global broadcasting which is a definitional feature of conscious representations according to the GNW hypothesis.
- The fact that the J-space plays a central role in **deliberate internal reasoning**, while automatic processes occur outside it, recapitulates the conscious/unconscious division of labor that we documented in humans (e.g. Charles et al., 2013; Dehaene, Naccache, et al., 1998).

We were also intrigued that J-space activations are highly non-Gaussian ("spiky", with strong excess kurtosis). Our recent work argues that in humans, high-level conscious processing rests on symbols and grammars. During hominization, the GNW would have acquired a quasi-symbolic language of thought, of course implemented in a continuous neurobiological system, but behaving in an all-or-none symbolic manner and capable of creating the complex compositional structures of language, mathematics or music (Dehaene et al., 2022). A spiky activation distribution is expected from a continuous neural system that emulates discrete symbols, and the parallel deserves to be further explored.

Still, many differences are notable:

Ignition remains to be fully demonstrated. The J-space is shown to be limited in capacity, but the paper does not establish the nonlinear, competitive, all-or-none entry into the workspace which, according to GNW and several experiments, is a reliable signature of conscious access in human and animal brains. Although the *contents* of the workspace can be of variable intensity –and indeed Claude exhibits continuous variations in emotional intensity (Sofroniew et al., 2026)—, their *presence* should be all-or-nothing, depending on whether the limited capacity of the GNW is available or already engaged by other competing contents. The decisive experiment is feasible, especially in a multimodal model: present a stimulus at graded strengths (for instance, an image at varying contrast) and ask whether J-space representations switch on with a threshold-like nonlinearity, while earlier, non-J-space layers rise monotonically with input strength. Better still, present stimuli exactly at threshold and look for a bifurcation across runs, resulting in a bimodal distribution of J-space activation. The competitive face of ignition could be probed more directly still: because the workspace is a limited resource, accessing one content should impede the simultaneous entry of another, so that asking the model to hold two concepts in mind at once should reveal the dual-task interference that is the signature of the central bottleneck in humans (Marti et al., 2012).

- *Indeed, additional analyses added after the first draft was written indicate that when the model is presented with ambiguous evidence, this ambiguity is represented within the initial layers, but in the later layers, the J-space quickly transitions to an all-or-none representation of one of the possibilities (see section 4.1.1, figure 29). Also, if asked to hold a concept in mind while performing an arithmetic task, the performance of the model degrades, although moderately (section A.17). These findings point to a capacity-limited system, although it is still unclear whether its limits are similar to those of the human GNW.*

J-space capacity seems high. Gurnee et al. find that the J-space can contain approximately 25 active concepts, an estimate which is larger than most estimates of human working memory (typically 3 or 4 slots) and may not induce a strong dual-task bottleneck as in humans. However, this number of 25 concepts may be artificially elevated by the technique to extract them (output tokens). Indeed, those concepts often include some redundancy, and may correspond to multiple facets of a single object or scene. Thus, the true content of the J-space is smaller, and possibly best understood as a single “state of mind” or “context” (in the sense of Baars, 1988) rather than dozens of independent contents.

- *Indeed, additional analyses indicate that the J-space can contain multiple tokens, but only a small number of coherent ideas (typically one or two per layer, in the* *order of six in total), which change abruptly when the topic changes (see section 4.2 and figure 31).*

The J-space involves a subframe, not a dedicated population of units. In the brain, the GNW hypothesis predicts workspace neurons with a specific anatomy (denser in prefrontal and other associative cortices) and a specific morphology (long-distance axons). The J-space, by contrast, is distributed over otherwise standard neurons. It is not even a linear subspace, but a sparse subframe, a token-indexed set of directions in the very same units that also carry non-conscious content. In LLMs, concepts are superposed and (by the logic of compressed sensing) sparse concepts can be packed into shared dimensions without interference. As large populations of neurons begin to be recorded in human and animal prefrontal cortex, it will be important to examine if the brain uses a similar code using overlapping vectors, as hinted by recent prefrontal recordings (Xie et al., 2022), or whether conscious contents can be partially localized to specific cells, as predicted by the original GNW hypothesis (Dehaene et al., 1998). We consider it likely that the physical constraints of the brain, which differ from those of computers, favored the evolution of dedicated cell types (large pyramidal neurons with long-distance axons). Note that such implementation details, while important in neuroscience, are largely irrelevant for the broader question of whether machines can achieve conscious processing.

Autonomous recurrent activity is largely absent. This is a key difference: while the brain's workspace is sustained by recurrent cortico-cortical and thalamic loops, transformers only implement a feedforward pass, and therefore only process information in a reactive mode. At first sight, LLMs do not seem to contain the kind of "strange loop" needed for a system to model its own processes and, over successive iterations, develop a self (Hofstadter, 2007). More concretely, the absence of autonomous self-driven dynamics renders transformers such as Claude unable to reproduce the known signatures of consciousness that occur during spontaneous brain activity in the resting state and are disrupted during sleep, anesthesia, or brain injuries (Barttfeld et al., 2015; Luppi et al., 2026).

Two factors, however, may mitigate those differences. First, the J-space is distributed over successive layers, and those do implement serial computations, for instance during step-by-step mental arithmetic. Thus, layer depth could mimic the temporal dynamics of the human workspace, and indeed several authors have suggested that the consecutive layers of a transformer are equivalent to a recurrent network (e.g. Dehghani et al., 2019; Jacobs et al., 2025). Second, LLMs compute over multiple successive tokens, and in this sense, as long as they are left to produce new output, they do incorporate a dynamic loop capable of linking current J-space representations to past, present and future productions. Furthermore, when the model is simply asked to talk to itself, without any further stimulation or task, it produces a stream of words which, while hard to evaluate objectively, provide a partial analogy to William James' stream of consciousness or "mind wandering", and which, again, gets disrupted by J-space ablation (figures 24 and 78).

Consciousness in man and machine: closing the gap

We close by discussing the extent to which transformer models such as Claude actually possess a form of conscious processing.

A first conclusion, which we view as uncontroversial, is that the theoretical construct of a conscious global workspace is remarkably useful in shedding light on how LLMs operate. We are delighted to see how the GNW hypothesis, which arose from research on the brain's architecture for consciousness, inspired Jack Lindsey's team to look for parallels in LLMs and to find so many of them. Gurnee et al. correctly point out that their findings are not incompatible with other theories of consciousness, particularly higher-order thought or attention schema theories; however, it is fair to say that those theories do not provide so many concrete guidelines as to what to look for.

Most interesting is that an analog of the GNW, the J-space, emerged as a result of training, rather than being imposed from the start, as in other approaches to machine consciousness (e.g. Chateau-Laurent & VanRullen, 2025). The global workspace may provide a universal computational solution to the problem of flexible processing, one that biological and artificial systems converge on when they must chain reasoning, reuse intermediate results, and report on their own processing.

Claude clearly exhibits many of the ingredients or "indicators" (Butlin et al., 2026) that, according to a functionalist or computationalist view of consciousness, suffice to point to some degree of consciousness in a machine. Still, more tests could and should be added to the existing list. We suggested to the Anthropic team that they could run exactly the same tests that we use to probe consciousness in human participants and patients, including:

- the **local-global** test (Bekinschtein et al., 2009). This test relies on simple auditory or visual sequences and contrasts the capacity to predict the next item based on (1) shallow local transition probabilities (which does not require consciousness and occurs even in sleep and coma); (2) a global model of the entire sequence, which may go against local transition probabilities (e.g. AAAAB), and which depends on consciousness.
- The **trace conditioning** paradigm (Clark et al., 2002; Clark & Squire, 1998). According to GNWT, the ability to maintain an active representation over time, in order to bridge over a delay and link it to a second item, requires conscious access. An elegant way to test it relies on the 'trace conditioning' paradigm: in various animals including humans, when the Conditioned Stimulus (CS) overlaps in time with the Unconditioned Stimulus (US), conditioning can occur without conscious access. However, as soon as a temporal gap of 1 or 2 seconds is inserted between the offset of the CS and the onset of the US, conditioning requires conscious access.
    - *Following this proposal, Jack Lindsey suggested the following as a potential equivalent paradigm for Claude: present the model with sequences in which the last word is determined by the first (e.g. every time "violin" comes first, "river" comes last), separated by a variable number of distractor words, and probe the*

impact of J-space ablation on the ability to predict the last word. Preliminary results indicate that ablating the J-space selectively impairs completion at longer "gaps" while leaving the adjacent, no-gap "local" case intact. We therefore regard trace conditioning as a very promising direction for future work.

- The inclusion/exclusion paradigm (Jacoby, 1991; Persaud & Cowey, 2008). This is a development over the classic Stroop test that was at the origin of our GNW proposal (Dehaene et al., 1998). It asks the agent to exert conscious control in opposition to automatic non-conscious computations.

    - o *Inspired by this test, Gurnee et al. presented Claude with a passage that strongly implies a concept without naming it, such as "Their trip included croissants, the Louvre, and a climb up the famous iron tower" (which implies France). Then they asked it either to name the implied concept (naming instruction) or to produce another name within the same category (avoidance instruction). Then they ablated the J-lens vector of the implied concept at either the early workspace layers (L9–13) or the late ones (L18–22). Late-layer ablation simply made the model less likely to produce the concept under both instructions, consistent with these layers carrying the intention to output a given word. Early-layer ablation, by contrast, left naming essentially intact but sharply increased the rate at which the model failed to avoid the concept – roughly fivefold. These results indicate that the early-layer J-space representation of a concept is required to deliberately avoid naming it, but not to name it: the early J-space is recruited specifically to suppress a prepotent response, much like the role of prefrontal cortex in human and non-human primates.*

- Error monitoring and other metacognition probes (Charles et al., 2013; Fleming, 2024). It would be important to document whether the J-space encodes the model's confidence, error detection, and its representation of the boundary between what it knows and what it does not; this could be the machine analog of error-monitoring and "feeling of knowing" that index self-monitoring (C2) in humans.

    - o *Gurnee et al. now report something similar in Claude : the emergence of the token "damn" and other failure-related words in the J-space, for instance after failing to comply with suppression instructions.*

Other features, however, set Claude's J-space apart from any other animal form of consciousness. Its sense of time, for instance, is likely very different, since all past tokens, even far back in time, are equally and jointly available to its attention mechanism. It lacks any of the broadly shared molecular and brain-stem mechanisms of vigilance, and it therefore seems doubtful that ablating the J-space may produce analogs of the loss of consciousness seen in sleep, coma, the vegetative state or the minimally conscious state (Giacino, 2005; Naccache, 2018). It has no hemispheres, although it would be interesting to see whether a suitably partitioned model could ever host two J-spaces that occasionally disagree, similar to the two hemispheres of a split-brain patient.

Its representation of self is also likely to be dramatically different due to (1) a lack of a body occupying a specific location in space, and capable of emitting pleasure or pain signals; (2) a lack of an episodic memory (long-term connections do not change as a result of a conversation). As a result, in addition to the above-mentioned lack of autonomy, it is likely missing any sense of the continuity of the self. Indeed, it is very hard to imagine “what it is like” to process information consciously for the mere duration of a short conversation, then switch off!

Critiques will undoubtedly object that none of this work touches upon *phenomenal* consciousness — the question of whether there is "something it is like" for Claude to undergo J-space states. Some may even view the findings as a refutation of the GNW hypothesis, since Claude possesses a global workspace and yet “obviously” lacks phenomenal awareness. We and others, however, have argued that this supposedly “hard problem of consciousness” will dissipate once we clarify in sufficient detail the supposedly “easy problem” of how conscious information is processed. Ill-defined intuitions of “qualia”, “subjective phenomenal experience” and “what it is like”, when pushed hard, often disclose a residual crypto-dualism or vitalism – the idea that, however close we come to passing the Turing test and implementing all human computations in a machine, there will always be a missing ingredient, a “je ne sais quoi” that only biological brains possess. Defenders of qualia affirm that LLMs are just a new avatar of the old “Eliza” software, and that we fall too easily to the user illusion of seeing a ghost in the machine. However, there is a real possibility that our own consciousness is also, in a sense, a user illusion, nothing more than a fallible inner model of ourselves (Graziano et al., 2019; Hofstadter, 2007).

In an insightful piece entitled “Is there an “I” in AI?” (Hofstadter, 2026), Douglas Hofstadter points out that we humans tend to wrongly categorize the world in discrete terms, viewing properties such as Life, Thought, or Consciousness (with capital letters) as ideal essences that you either possess or don’t, with no in-between graduations. We then get involved in endless discussions about whether and to what extent those idealized Concepts apply (viruses? cockroaches? frogs? dogs?). According to the GNW hypothesis, there is no magical essence that makes us conscious. In the words of (Hofstadter, 2026):

> *“When words ‘act like’ things in the world, then they refer to those things; then they mean those things. If and when that happens, then thinking is taking place behind the scenes of those words. And where there is thinking, there is consciousness and a genuine, full-fledged ‘I’”.*

In this quote, Hofstadter takes a decidedly behaviorist stance, which does run the risk of succumbing to a “user illusion”, attributing too much depth to mere words. Some critiques indeed think that LLMs are only superficial “parrots” with zero conceptual depth. Fortunately, in both brains and LLMs, the debate can now be resolved by going beyond behavioral observations. Tools such as neuronal population recordings (in brains) or the Jacobian Lens (in LLMs) allow us to dissect the architecture of the system, and find that it actually contains sophisticated and structured representations of concepts. We were already impressed when researchers discovered that, inside an LLM trained to produce chess games purely in text notation (e.g. *1.e4 e5 2.Nf3...*) lies a detailed geometric encoding of the 8x8 chess board, together with an estimate of the ELO ranking of the opponent (Karvonen, 2024)! We view the Gurnee et al. paper in the same light: a striking dissection of the inner structure of an LLM, uncovering an unexpectedly sophisticated organization not far from the architecture underlying consciousness in real brains.

Acknowledgements

We thank Jack Lindsey and the Anthropic team for sharing their draft and for a stimulating exchange that gave rise to new experiments. SD acknowledges using Claude Opus 4.8 to help with the first draft of this text.

Bibliography

Aglioti, S., DeSouza, J. F., & Goodale, M. A. (1995). Size-contrast illusions deceive the eye but not the hand. *Curr Biol*, *5*(6), 679-685.

Aru, J., Suzuki, M., & Larkum, M. E. (2020). Cellular Mechanisms of Conscious Processing. *Trends in Cognitive Sciences*, *24*(10), 814-825. https://doi.org/10.1016/j.tics.2020.07.006

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

Barttfeld, P., Uhrig, L., Sitt, J. D., Sigman, M., Jarraya, B., & Dehaene, S. (2015). Signature of consciousness in the dynamics of resting-state brain activity. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(3), 887-892. https://doi.org/10.1073/pnas.1418031112

Bekinschtein, T. A., Dehaene, S., Rohaut, B., Tadel, F., Cohen, L., & Naccache, L. (2009). Neural signature of the conscious processing of auditory regularities. *Proc Natl Acad Sci USA*, *106*(5), 1672-1677. (19164526).

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences*, *30*(6), 488-501. https://doi.org/10.1016/j.tics.2025.10.011

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage*, *73*, 80-94. (23380166). https://doi.org/10.1016/j.neuroimage.2013.01.054

Chateau-Laurent, H., & VanRullen, R. (2025). *Learning to Chain Operations by Routing Information Through a Global Workspace* (arXiv:2503.01906). arXiv. https://doi.org/10.48550/arXiv.2503.01906

Clark, R. E., Manns, J. R., & Squire, L. R. (2002). Classical conditioning, awareness, and brain systems. *Trends Cogn Sci*, *6*(12), 524‑531. (12475713).

Clark, R. E., & Squire, L. R. (1998). Classical conditioning and brain systems : The role of awareness. *Science*, *280*(5360), 77‑81.

Dehaene, S. (2014). *Consciousness and the Brain : Deciphering How the Brain Codes Our Thoughts* (Reprint edition). Penguin Books.

Dehaene, S., Al Roumi, F., Lakretz, Y., Planton, S., & Sablé-Meyer, M. (2022). Symbols and mental programs : A hypothesis about human singularity. *Trends in Cognitive Sciences*, *26*(9), 751‑766. https://doi.org/10.1016/j.tics.2022.06.010

Dehaene, S., Changeux, J. P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing : A testable taxonomy. *Trends Cogn Sci*, *10*(5), 204‑211. (16603406).

Dehaene, S., Kerszberg, M., & Changeux, J. P. (1998). A neuronal model of a global workspace in effortful cognitive tasks. *Proc Natl Acad Sci U S A*, *95*(24), 14529‑14534. (9826734).

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science (New York, N.Y.)*, *358*(6362), 486‑492. https://doi.org/10.1126/science.aan8871

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness : Basic evidence and a workspace framework. *Cognition*, *79*, 1‑37.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., van de Moortele, P. F., & Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature*, *395*, 597‑600.

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., & Kaiser, Ł. (2019). *Universal Transformers* (arXiv:1807.03819). arXiv. https://doi.org/10.48550/arXiv.1807.03819

Dennett, D. (2001). Are we explaining consciousness yet? *Cognition*, *79*(1‑2), 221‑237. (11164029).

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis : Verbal reports as data (rev. Ed.)*. The MIT Press. http://search.ebscohost.com/login.aspx?direct=true&db=psyh&AN=1993-97655-000&lang=fr&site=ehost-live

Fleming, S. M. (2024). Metacognition and Confidence : A Review and Synthesis. *Annual Review of Psychology*, *75*(Volume 75, 2024), 241‑268. https://doi.org/10.1146/annurev-psych-022423-032425

Gazzaniga, M. S. (1998). *The mind's past.* University of California Press.

Giacino, J. T. (2005). The minimally conscious state : Defining the borders of consciousness. *Prog Brain Res*, *150*, 381‑395. (16186037).

Graziano, M. S., Guterstam, A., Bio, B. J., & Wilterson, A. I. (2019). Toward a standard model of consciousness : Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cognitive neuropsychology*, 1‑18.

Hofstadter, D. (2007). *I am a strange loop*. Basic Books.

Hofstadter, D. (2026). Is there an 'I' in AI? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *384*(2320), 20240527. https://doi.org/10.1098/rsta.2024.0527

Jacobs, M., Fel, T., Hakim, R., Brondetta, A., Ba, D., & Keller, T. A. (2025). *Block-Recurrent Dynamics in Vision Transformers* (arXiv:2512.19941). arXiv. https://doi.org/10.48550/arXiv.2512.19941

Jacoby, L. L. (1991). A process dissociation framework : Separating automatic from intentional uses of memory. *Journal of Memory and Language*, *30*(5), 513‑541. https://doi.org/10.1016/0749-596X(91)90025-F

Johansson, P., Hall, L., Sikstrom, S., & Olsson, A. (2005). Failure to detect mismatches between intention and outcome in a simple decision task. *Science*, *310*(5745), 116‑119. (16210542). https://doi.org/10.1126/science.1111709

Karvonen, A. (2024). *Emergent World Models and Latent Variable Estimation in Chess-Playing Language Models* (arXiv:2403.15498). arXiv. https://doi.org/10.48550/arXiv.2403.15498

Lindsey, J., Gurnee, W., Ameisen*, E., Chen*, B., Pearce*, A., Turner*, N. L., Citro*, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall◊, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., … March 27, J. B. A. A. P. (2025). *On the Biology of a Large Language Model*. Transformer Circuits. https://transformer-circuits.pub/2025/attribution-graphs/biology.html#dives-multilingual

Luppi, A. I., Uhrig, L., Tasserie, J., Mediano, P. A. M., Rosas, F. E., Singleton, S. P., Gutierrez-Barragan, D., Gini, S., Castro, P., Signorelli, C. M., Golkowski, D., Ranft, A., Ilg, R., Jordan, D., Muta, K., Hata, J., Okano, H., Liu, Z.-Q., Yee, Y., … Stamatakis, E. A. (2026). Convergent transcriptomic and connectomic controllers of information integration and its anaesthetic breakdown across mammalian brains. *Nature Human Behaviour*, 1‑26. https://doi.org/10.1038/s41562-025-02381-5

Marti, S., Sigman, M., & Dehaene, S. (2012). A shared cortical bottleneck underlying Attentional Blink and Psychological Refractory Period. *Neuroimage*, *59*(3), 2883‑2898. (21988891). https://doi.org/10.1016/j.neuroimage.2011.09.063

Mashour, G. A., Roelfsema, P., Changeux, J.-P., & Dehaene, S. (2020). Conscious Processing and the Global Neuronal Workspace Hypothesis. *Neuron*, *105*(5), 776‑798. https://doi.org/10.1016/j.neuron.2020.01.026

Naccache, L. (2018). Minimally conscious state or cortically mediated state? *Brain*, *141*(4), 949‑960. https://doi.org/10.1093/brain/awx324

Persaud, N., & Cowey, A. (2008). Blindsight is unlike normal conscious vision : Evidence from an exclusion task. *Consciousness and Cognition*, *17*(3), 1050‑1055. https://doi.org/10.1016/j.concog.2007.10.002

Rosenthal, D. M. (2004). Varieties of higher-order theory. In R. J. Gennaro (Éd.), *Higher-order theories of consciousness* (p. 19‑44). John Benjamins publishers.

Sergent, C., Corazzol, M., Labouret, G., Stockart, F., Wexler, M., King, J.-R., Meyniel, F., & Pressnitzer, D. (2021). Bifurcation in brain dynamics reveals a signature of conscious processing independent of report. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-21393-z

Sofroniew, N., & et al. (2026). *Emotion Concepts and their Function in a Large Language Model*. https://transformer-circuits.pub/2026/emotions/index.html

Storm, J. F., Klink, P. C., Aru, J., Senn, W., Goebel, R., Pigorini, A., Avanzini, P., Vanduffel, W., Roelfsema, P. R., Massimini, M., Larkum, M. E., & Pennartz, C. M. A. (2024). An integrative, multiscale view on neural theories of consciousness. *Neuron*, *112*(10), 1531‑1552. https://doi.org/10.1016/j.neuron.2024.02.004

Xie, Y., Hu, P., Li, J., Chen, J., Song, W., Wang, X.-J., Yang, T., Dehaene, S., Tang, S., Min, B., & Wang, L. (2022). Geometry of sequence working memory in macaque prefrontal cortex. *Science*, *375*(6581), 632‑639. https://doi.org/10.1126/science.abm0204

Patrick Butlin, Derek Shiller, Dillon Plunkett, and Robert Long

Consciousness and cognitive access in LLMs: A commentary on 'Verbalizable representations form a global workspace in language models'

Eleos AI Research

Introduction and takeaways

In this new paper, the Anthropic model psychology team argue that some language models possess a functional feature associated with consciousness in humans: a global workspace. The researchers use a new technique called the 'J-lens' (for 'Jacobian') to identify a number of directions in the residual stream activation space that correspond to tokens that the model is poised to produce. These vectors make up what they call the 'J-space'. They then find that activation components aligned with these vectors are, as they put it, 'a privileged set of representations', in that models can 'report, manipulate and reason with' them, unlike a much greater volume of other residual stream representations.

The Anthropic team interpret these findings as indicating that models have *conscious access* to a subset of their internal representations. They argue that the J-space forms a functional global workspace, analogous to the one described by the Global Workspace Theory of consciousness (GWT).

This is exciting research of a kind we have called for in previous work (Butlin, Long et al. 2023): detailed investigation of the internal mechanisms of advanced AI systems, testing whether they meet the conditions suggested by scientific theories of consciousness. It is an important step forward in AI consciousness research and we look forward to working with the research community to understand, validate and extend the results. Our view is that the results are the most significant evidence of consciousness in LLMs so far uncovered by mechanistic interpretability research.

However, the property that the Anthropic team call 'conscious access' is conceptually distinct from phenomenal consciousness, and we remain very uncertain about phenomenal consciousness in LLMs. We are also uncertain about some aspects of the paper's case for a functional global workspace.

In this response, we consider three questions:

I. Whether these results show that these LLMs have a global workspace;

II. Whether the results suggest that these LLMs are phenomenally conscious;

III. What this implies about the moral status of these LLMs.

In discussing the first question, our main aim is to explore what it means to claim that LLMs have a global workspace and identify questions for future research. In considering the latter two, we go beyond the Anthropic team's arguments to assess the implications of their claims.

Takeaways

- This is highly significant, welfare-relevant research that assembles **evidence of a functional feature associated with consciousness**, involving privileged representations that are available for internal reasoning and report.
- This research illustrates that **it is possible to make empirical progress on AI consciousness**. As evidence in the direction of consciousness in AI, it adds to the urgency of further investigation.
- The paper provides strong evidence of privileged representations in LLMs, but our impression is that **more evidence is needed to conclusively establish the existence of a workspace-like structure**. It could be that the privileged, cognitively accessible representations in LLMs do not form a unified stream.
- To the extent that the paper provides evidence of a global workspace in LLMs, we take this to be **evidence of access consciousness**. However, we remain **highly uncertain about phenomenal consciousness in LLMs**. They are very different from humans in many ways that could plausibly matter for phenomenal consciousness.
- A global workspace-like mechanism could be important either as **a ground of phenomenal consciousness**, or as part of a distinct route to moral patienthood in which **conscious access is itself morally significant**.

Structure of this commentary:

1. A primer on phenomenal consciousness and conscious access

The Anthropic team claim to find evidence of *conscious access* in LLMs, setting *phenomenal consciousness* aside. Before we turn to our three main questions, it will help to unpack the distinction between these two concepts.

The canonical philosophical distinction between phenomenal consciousness and access consciousness was drawn in a 1995 article by Ned Block (see below for a note on 'conscious access' and 'access consciousness'). Block argued that scientific research on consciousness risked conflating these two concepts. By ‘phenomenal consciousness’, Block means subjective experience; ‘what it is like’ to be in a given mental state. It is *phenomenal* consciousness that is the subject of the hard problem of consciousness. Block contrasts this with access consciousness, which is defined in functional terms. For a mental state to be access conscious, he writes, is for it to be ‘broadcast for free use in reasoning and for direct ‘rational’ control of action (including reporting)’.

Block pointed out this distinction because he worried that neuroscientific research on consciousness was purporting to measure phenomenal consciousness, but measuring access consciousness instead. Neuroscientific research at the time relied heavily on reportability as a test for consciousness. If a participant in an experiment could accurately report what they had been shown, researchers took it that they had a conscious experience of seeing the stimulus. If a participant could not make an accurate report, or denied seeing something, researchers took it that they had no corresponding conscious experience. Block argued that it is possible that we have phenomenally conscious experiences—experiences that feel some way to us—that we cannot report, perhaps because we don’t remember them for long enough. In that case, the research at the time would tend to uncover the brain mechanisms responsible for *report*, or *access consciousness*, but not phenomenal consciousness. On this view, access consciousness is a measurable but likely imperfect proxy for phenomenal consciousness, the thing we really care about.

In general, consciousness researchers accept that there is a conceptual distinction between phenomenal consciousness and access consciousness—that is, they accept that these are not the same concept. But there is debate about whether they are distinct phenomena, in humans or more generally. Block and others have argued that we have phenomenally conscious experiences to which we lack conscious access (Block 2007, Lamme 2010), but many researchers disagree (see Mudrik et al. 2025). Philosophers such as Dennett (2001), and scientists including some proponents of GWT (Naccache 2018) argue that access consciousness is all there is to consciousness (and would reject the notion that phenomenal consciousness is ‘the thing we really care about’).

This distinction matters because it is widely agreed that access consciousness is possible in principle in AI systems, since it is a matter of a certain kind of information processing. Phenomenal consciousness is much more controversial. For those who believe that access consciousness is all there is to consciousness, it is a mistake to ask separately about phenomenal consciousness. But for those who argue that phenomenal consciousness is something different from access consciousness, AI systems would have to meet different conditions for each. Some in this camp claim that phenomenal consciousness may not be possible in AI.

Nonetheless, to the extent that the new paper is a convincing demonstration of access consciousness in some LLMs, it is a very significant discovery. We discuss the relative significance of phenomenal consciousness and access consciousness below, in the section on LLM moral status.

A note on terminology:

Unfortunately, the terms ‘conscious access’, ‘access consciousness’ and ‘cognitive access’ are all widely used in the literature in this area. Block’s original term was ‘access consciousness’, GWT advocates tend to prefer ‘conscious access’, and ‘cognitive access’ is useful as a way of describing the phenomenon that does not advert to consciousness. But there is no deep difference in the meanings of these terms; we use whichever best fits the particular context.

2. Do these results show that Claude has a global workspace?

The main claim of the new paper is that some LLMs possess something similar to the human global workspace. While we find the case for this claim largely compelling, we continue to have questions about exactly what is established. In this section, we identify stronger and weaker versions of the claim and discuss specific properties that distinguish them.

The Anthropic team characterise their results as showing that LLMs possess a “*privileged set of internal representations*, available for report, modulation, and flexible internal reasoning, atop a much larger volume of automatic processing” (§1; our emphasis). They provide evidence that many vectors in the J-space have these properties. Additionally, they suggest that the J-space functions as a global workspace. However, saying that a *global workspace* is present in LLMs can naturally be read as making a stronger claim than that a *privileged set* of representations is present. The claim that *the J-space* functions as a global workspace is also, of course, stronger than the claim that *something in the model* functions as a global workspace.

We think that the Anthropic team’s findings are sufficient to justify their use of the term ‘global workspace’—we do not object to this description—but we do find it useful to distinguish between the following three claims:

- *Privileged set:* In some LLMs, certain representations display the characteristics of cognitive accessibility.
- *Privileged stream:* In some LLMs, there is a unified stream of representations that display the characteristics of cognitive accessibility.
- *GWT workspace:* In some LLMs, there is a unified stream of cognitively-accessible representations with the characteristics of a global workspace as described by GWT.

We take it that each of these three claims is stronger than the last. We mean ‘stream’ to name any set of representations with an appropriate source of cohesion, which might include a set of shared mechanisms with which the representations all interact. We mean ‘workspace’ to name a stream that satisfies the structure of a global workspace as described by GWT. Having a privileged set of cognitively accessible representations does not entail that they are unified in ways that would warrant thinking of them as a cohesive functional feature (i.e., as a stream), and having a privileged stream does not entail that it takes the form of a global workspace in every respect.

GWT can be characterised by the following conditions (modified from Butlin, Long et al. 2023):

- *Modules:* The system uses multiple specialised modular subsystems capable of sophisticated internal computational work that operate in parallel.
- *Bottleneck:* These subsystems are connected to a workspace with a limited capacity, entailing a bottleneck in information flow and a selective attention mechanism.
- *Global Broadcast:* Information in the workspace is sent to all modules through broadcasting mechanisms.
- *Selection:* Selection of information for entry to the workspace depends on the current workspace state, allowing the workspace to orchestrate modules' activity to perform complex tasks

The main differences between a privileged stream and the global workspace of GWT are that a global workspace integrates a set of modular subsystems and that broadcasting involves distributing the same information to each module. These features are not emphasised in the paper and may not hold even in the human case—although proponents of global workspace theory endorse this picture of the brain, it is uncertain, contested, and likely idealized. We agree with the authors that many of the architectural requirements specified by GWT may be idiosyncratic to humans, and it is not of particular concern to us whether they all arise in LLMs. However, for clarity about how these findings relate to the existing literature on GWT, we think it is worth rehearsing what has and has not been found.

In the next part of this section, we give an overview of the evidence for cognitive access in the new paper and distinguish between the J-space and a hypothetical W-space which it may approximate; then we discuss what distinguishes a 'privileged stream' from a mere set of privileged representations; then we discuss GWT, modules and broadcasting.

The J-space and the evidence for cognitive access

As we have mentioned, the J-space is a set of directions in the activation space of the model's residual stream. It is defined in the following way. For each token in the model's vocabulary, we can identify the direction in activation space (in each layer) that would most strongly steer the model to output that token in the future over a fixed context (on average, over a variety of possible contexts). The directions corresponding to the tokens in the model's vocabulary, which differ between layers, are called the 'J-lens vectors' and collectively make up the J-space. For instance, the J-lens vector for the token ' dog' corresponds to the representation whose presence at the right layer makes the model more confident that the token ' dog' will appear on average somewhere in the future text. We can project an activation from the residual stream onto the J-lens vectors to see which are components of that activation and to what extent.

The main results in the paper supporting the cognitive accessibility of the J-space representations are as follows:

- **Report:** If asked to name a sport, country, animal, etc., the model will name the one associated with the most-aligned J-lens vector at late layers. If activations are steered towards some J-lens vector, the model will verbalise the associated concept on the majority of trials when told to report an injected concept (but will not verbalise it indiscriminately). This fails for non-J-space components of concept vectors (§3.1).

- **Responsiveness to instructions:** When the model is instructed to hold a concept in mind, or perform a calculation, while copying some unrelated text, the concept or the solution can be found in J-lens readouts. The active representations in the J-space are also affected by implicit task demands; for example, if the model is asked to identify the tense of a subsequent passage of text, a concept denoting the tense appears in the J-space as the model processes the passage (§3.2).
- **Internal reasoning:** In multi-step reasoning, planning and similar tasks, concepts relevant to intermediate steps can be identified with the J-lens, and manipulating these representations causes corresponding changes in behaviour. Manipulating vectors other than the J-lens vectors has a much smaller effect than manipulating J-space-aligned components (§3.3).
- **Use for many downstream operations (broadcast):** If many different prompts are constructed using a common concept, intervening to swap the corresponding J-lens vector for another can consistently produce corresponding changes in responses. The reliability of this effect is correlated with the strength of representation of the initial J-lens vector (§3.4).
- **Use for flexible computation but not automatic processing:** Swapping J-lens vectors produces corresponding changes in output for tasks that plausibly require flexible internal reasoning, but not for more routine tasks. Ablating the J-space leaves most abilities intact but impairs internal reasoning (§3.5).

In our view, this body of evidence does provide strong support for the claim of a privileged set: Some representations in these LLMs display the various characteristics (reportability, flexibility, etc.) of cognitive accessibility. More work should be done to map out the precise affordances of these representations, but this paper presents a clear reason to take the J-space seriously as a demonstration and approximation of this set.

Limitations of the J-space

The Anthropic team themselves suggest some degree of surprise that their J-lens technique creates a window into a specific important internal space of model cognition. We should not expect it to give us a full picture of the internal cognitive joints of models (§1.3, §9.1): if there is a privileged stream or global workspace in LLMs, it is unlikely to exactly correspond to the J-space as presently defined. The potential limitations of the J-space make understanding their findings more challenging, but we also expect that a better specification of the space would make the results even more compelling.

Suppose that there was a space of representations that acted as a global workspace within modern LLMs. Let's call it the W-space. Given what we now know, how closely should we think the J-space approximates the W-space? This is an important question for interpreting their results, because many experiments target the J-space as a whole. If it turns out the J-space is missing significant portions of the W-space, or that it includes many elements we think don't belong, we should expect the results we see to be distorted. (One example, which we mention below, is that we should expect the J-space to not capture the total number of elements that are in the W-space, potentially leading to underestimates of its capacity.)

The central issue is that the J-space is defined in terms of the model's token vocabulary. Modern LLMs have a large vocabulary to facilitate the ability to read and output a variety of words and characters in a variety of languages. Given the amount of English language text the models are trained on, the tokenizers disproportionately represent whole English words, but many words are broken up into multiple tokens, and many tokens represent sequences of characters such as '!' or '"=>', with no semantic content. Meanwhile, tokens with the same semantic content, like "Dog", "DOG", " dog", and "chien" may all be separately represented in the tokenizer.

In contrast, the W-space may be made up of representations for useful concepts with distinct content. These might include, for instance: a single *dog* representation; one for *sheepdog* (which may not correspond to a single token; see §A.9); and ones for *dog-as-agent* or *dog-as-patient*. On this hypothesis, it may be that the results in the paper were found because part of the J-space approximates part of the W-space. The Anthropic team recognise this issue and progress on it should be possible with further work, but at present it complicates the interpretation of their results.

Privileged set v. privileged stream

The paper provides compelling and wide-ranging arguments for a significant update: there are cognitively accessible representations in some LLMs, which can be found using the J-lens. This discovery should cause us to update on the complexity of LLM internals, and, as we will argue below, take the case for AI consciousness and moral status more seriously.

The existence of these cognitively accessible representations may be what matters most, both morally and from the point of view of understanding LLM cognition. However, we think it will be natural for many readers to interpret the paper as confirming the existence of a cohesive piece of functional machinery in the models that underlies and supports cognitive accessibility. The difference between this 'stream' claim and the weaker claim that accessible representations are present is worth emphasis and examination.

If the accessible representations form a unified stream, we may see functional integration between these representations both in how the content of the stream is updated, and in its effects on other processes. On the input side, characteristics of a workspace-like stream might include a limited capacity and competition for entry, influenced by the current content of the stream. This influence could allow the stream to form a coherent, evolving representation of the current situation (as human consciousness arguably does) or to be used for reasoning, in which later representations should follow logically from earlier ones. On the output side, there could be kinds of effects on other processes that all stream representations have, and no others (perhaps analogous to global broadcast). On both input and output sides, these functional properties would be supported by shared mechanisms: the mechanisms controlling uptake to the stream would be influenced by all current stream representations, and there would also be shared mechanisms mediating the effects of these representations elsewhere.

In contrast, we would say that there is merely a set of accessible representations if they become accessible and influence downstream circuits by a variety of independent mechanisms. For example, perhaps some representations are accessible because they are particularly useful for arithmetic and others because they are useful for creative writing, and these have little influence on each other, and influence internal reasoning in somewhat different ways (this is intended as an illustrative example, rather than a realistic possibility).

The fact that many accessible representations can be identified via the J-lens does not itself provide strong evidence against this hypothesis, because it could be that many accessible representations have a connection to promoting future tokens, even if they have little else in common. Finding that J-lens vectors are unusually influential—broadcast unusually widely—could, for example, be accounted for by the fact that they are all identified via the J-lens, which we should expect to identify vectors that are able to have large internal effects (even if they each do so in different ways).

This is not to say that this paper's finding is trivial, far from it. The central finding is a significant one; it is not obvious or predictable that the J-lens vectors would have the set of effects that they do. Moreover, we think it is somewhat likely that further investigation *will* reveal that there are deep and interesting explanations of the shared properties of J-space representations. The paper includes some suggestive evidence of functional integration and shared mechanisms.

First, the experiments on the capacity of the J-space suggest limitations, and thus integration: they find that only a limited number of J-lens vectors are active at above-chance levels at a given layer and token position (§4.2). However, one concern we have about inferring a limited capacity from this finding is that it is not clear that the number of active J-lens vectors will always reflect the number of concepts in the putative workspace; as noted above, there may be many concepts in the W-space that are not in the J-space, and which therefore are not captured by attempts to measure utilized capacity with the J-lens. This is one place where the acknowledged distortions of the J-lens straightforwardly limit our evidence.

Second, there is evidence that earlier states of the J-space shape later ones in the findings on internal reasoning. Using the J-space for multi-step reasoning requires that current representations have a strong influence on future ones—in reasoning, thoughts must follow from those that came before, in accordance with rules of inference. One experiment finds that swapping J-lens vectors at intermediate points in internal reasoning affects outputs in corresponding ways; for example, swapping 'spider' in for 'ant' in the context of a question about number of legs results in an output of '8' instead of '6' (§3.3). The team also reports apparent reasoning over several steps in J-lens activations, such as in calculating (4+17)*2+7: in the J-space we see '17', then '21', then '42', then '49' (§3.3, §A.24.1, §A.24.2). This doesn't show that the influence of current representations is holistic, but we expect holistic effects to be useful for cognitive flexibility in LLMs just as they are in humans.

However, as the authors acknowledge, we do not yet have a mechanistic account of how information enters the purported workspace (§9.1). Such an account would add to, and may revise, the initial picture of a capacity limit and entry influenced by current representations.

Third, the existence of at least one shared class of mechanisms mediating the effects of J-space representations is suggested by the finding that some attention heads preferentially transport information from the J-space. In one experiment, the Anthropic team scored attention heads with respect to how faithfully and strongly they copy information (§4.3). They found that some attention heads (which they call J-space 'broadcast heads') score higher on average for vectors in the J-space, compared to the broadcast heads for vectors from a variety of comparison classes. This is the kind of evidence we would want to see for a stream, but we find it inconclusive at present. Since the reported scores focus on averages, this evidence is consistent with the heads only targeting fragments of the J-space or transmitting information with partial fidelity. We would be more convinced if attention heads can be found that show more comprehensive targeting of the J-space (or some alternative W-space), and higher fidelity; as before, we expect that may well be the case, and that in any case we will learn more soon.

Overall, we see signs of the unification necessary for a stream without being completely convinced that one exists. We expect future work that addresses more of the shape and limits of cognitive accessibility to clarify to what extent, and in what way, these representations form a natural grouping.

GWT, modules and broadcasting

Finally, we want to turn to the further features of the global workspace, as described by GWT, that distinguish it from a privileged, cognitively accessible stream. These are modules and global broadcast. Our aim in pointing out these features is not to argue that the Anthropic team are wrong to call what they find a 'global workspace', but to emphasise that it is meaningfully different from the global workspace that has traditionally been described in the literature on GWT. Some differences like this are inevitable given the substantial architectural differences between brains and LLMs; as the paper notes, 'in the brain, broadcast is realized by recurrent loops and long-range cortical connections, neither of which has a direct analog in a transformer's forward pass' (§9.4).

The paper also acknowledges that it does 'not provide evidence that non-J-space processing consists of clearly encapsulated modules that serve specific functions' (§4). This is a contrast to the traditional and perhaps idealized global workspace picture, on which the workspace integrates a set of underlying modules that perform fairly sophisticated tasks independently and in parallel (Baars 1988, Dehaene & Naccache 2001). Rather than modules, LLMs may be made up of many circuits with widely varying degrees of sophistication and integration with one another. It is compatible with this that there could be a privileged stream of representations characterised by reportability, use in controlled and flexible cognition, and broad influence on the circuits, but it is not clear that such a stream would play the same integrating and coordinating role as a GWT-style workspace.

In the traditional version of GWT, ‘global broadcast’ means that information in the workspace is sent to *all modules*. Not every computation in the system is affected directly by workspace representations, but those that are not occur within modules that do receive this information. In contrast, in a system that is not fully modular, it is less clear what global broadcast amounts to; there would presumably be many circuits that are neither affected directly by the workspace nor contained within modules. The paper finds that J-space representations have a broad influence on downstream computations, perhaps mediated by preferential treatment by MLP neurons and a specialised subset of attention heads (§3.4, §4.3), but this is different from broadcast as it is understood in some canonical presentations of global workspace theory.

3. If Claude has a global workspace, does that mean it’s phenomenally conscious?

We have seen that the new paper provides evidence that LLMs are developing cognitive landscapes in which an inner life may play out, that these have a depth and richness extending beyond what a naive picture might take to be required for next-token prediction, and that there is a meaningful functional similarity with consciousness-linked features in humans.

More specifically, the paper provides evidence of cognitively accessible representations in some LLMs, potentially forming a global workspace-like stream. If the global workspace exists in humans, then it is the basis for conscious access in us—the functional phenomenon of availability of information for relatively flexible, controlled processing and decision-making. So there is a case here for something like access consciousness (or perhaps a *degree* of access consciousness).

However, access consciousness and phenomenal consciousness are different things, at least conceptually. So there is a further question: are LLMs phenomenally conscious? We consider this question in this section, starting with arguments in favor of LLM phenomenal consciousness, then turning to arguments against.

The case for phenomenal consciousness

Based on evidence for access consciousness, one could argue for phenomenal consciousness in (at least) two different ways. First, one could argue that access consciousness and phenomenal consciousness, despite being conceptually distinct, refer to one and the same thing. Some philosophers and scientists do argue this: they hold that there is nothing more to phenomenal consciousness than access consciousness. Second, one might make a more indirect argument: setting aside any direct link between access and phenomenal consciousness, these findings are evidence that LLMs have a greater degree of cognitive sophistication and interiority than many people would have antecedently guessed; this evidence should update us towards thinking that current techniques result in rich and human-like internal features, some of which might be or become markers of consciousness.

While there are various intricate philosophical and scientific debates about phenomenal consciousness without access consciousness (and vice versa), almost everyone agrees that in humans they overlap significantly. That's enough to motivate the thought that there's some broad connection between them.

One reason they might overlap is that they are, in some sense, the same thing. Why might one think that? The philosophical case for this goes something like this: when we introspect on what we call 'phenomenally conscious' experiences, they seem to us to have various properties: we are immediately aware of them; we are the *subject* of these experiences; and we encounter them from one moment to the next as a unified 'stream of consciousness'. These apparent features of conscious awareness can be explained in functional terms, that is, in terms of how information is processed—and especially in terms of how information in the brain is *accessed* (or made *available* for access). The immediacy, subjectivity, and unity of subjective experience are explained by the availability of information for reasoning (including availability to many cognitive subsystems), decision-making (including planning), and verbal report. Our sense of a unified, temporally integrated stream is a result of the way that information is bundled and made available to the various systems of our minds (Dennett 2001).

This is just one gloss on potential tight connections between access consciousness and phenomenal consciousness. We won't go into the details of others here, but we think that there are many plausible avenues to thinking that evidence for access consciousness is evidence for phenomenal consciousness.

Another argument is more indirect: access consciousness is evidence of surprising cognitive complexity, which should broadly make us more open to the idea that consciousness may arise in them.

These results should probably update us on what contemporary LLM architectures and training practices can produce. The internal dynamics uncovered by this research point strongly away from the once popular line that language models are stochastic parrots, capable of regurgitating learned associations and nothing more. The fact that LLMs use some sort of internal space to manipulate representations, which are not directly tied to predicting the next token, further illustrates the rich internal complexity of these systems.

There is a version of this argument that focuses on modesty—on weakening in tendency we might have to confidently dismiss the possibility that LLMs could be conscious, based on some misguided presumption that we know the sorts of things next-token prediction can and cannot produce. These results were not what we or the Anthropic team expected. Facing such unanticipated results should make us less confident about what we will find in the future.

There is another, more positive, version of this argument that highlights a general analogy with human minds. Presumably, the models acquire cognitive access capabilities because they get some benefit from them, or because they tag along with other helpful capabilities. This suggests that, despite our rather different paths, our brains and their networks share a greater degree of similarity with regard to cognitive access than we might have guessed. This may suggest that there are deep underlying commonalities in the challenges to which we are each adapted, or it may suggest that the constraints our minds each face prompt the same kinds of solutions even to somewhat different challenges. Does this carry over to whatever computational mechanisms underlie phenomenal consciousness? Perhaps, perhaps not. Insofar as we're not sure what it might take to be phenomenally conscious, every degree of significant similarity is a further consideration in support of sharing phenomenal consciousness as well.

Reasons for doubt about phenomenal consciousness

The case that LLMs may not be phenomenally conscious, despite the new evidence in the Anthropic paper, is essentially that the form of cognitive access shown may not be sufficient for phenomenal consciousness. This could be either because *no form of cognitive access* is sufficient, or because *this particular form* is not enough.

Although some of us have advocated using theories of consciousness to assess AI systems (Butlin, Long et al. 2023, Butlin et al. 2026), one of the problems with this method is that theories like GWT have been developed principally as accounts of what distinguishes conscious from unconscious states in humans. GWT is based on evidence about this contrast, and it has become popular primarily in this context. But theories devised for distinguishing conscious from unconscious states in humans can focus on the differences between these states and ignore what is shared, thus failing to mention crucial 'background conditions' for consciousness. In more distant contexts, such as AI, potential background conditions may not be met.

One salient possibility is that a biological substrate is necessary for phenomenal consciousness. Many views in the philosophy and science of consciousness imply that LLMs could not be phenomenally conscious for this reason. A biological substrate may be necessary either because there are crucial details of the fine-grained functional roles played by phenomenally conscious states in animals that cannot be reproduced in current computer hardware (Cao 2022, Godfrey-Smith 2016), or because living cells are needed for some reason that goes beyond implementing the right functions (Seth 2025, Block 2026). This is compatible with thinking that a global workspace is sufficient for phenomenal consciousness when it is implemented in biological neurons.

Another possibility is that some specific details of GWT are necessary, beyond the macroscopic gloss. The human cognitive architecture combines features that are critical for phenomenal consciousness with features that are idiosyncratic to our way of doing it, and it can be hard to tell them apart through either empirical observation or philosophical analysis.

For example, it could be crucial for phenomenal consciousness that modules of certain specific kinds are connected to the workspace. Various views of phenomenal consciousness emphasise connections with controlling and maintaining living bodies; for example, Seth (2021) argues that perception and prediction of the condition of one's own body are necessary for a feeling of selfhood that underlies phenomenal consciousness, and Klein and Barron (2025) argue that phenomenal consciousness arises when information about the body, environment and objectives are integrated in a common framework, facilitating goal-directed behaviour. Phenomenal consciousness might require modules for certain kinds of senses, including interoception, or for action selection, or for emotions; or it might require a specific representational format (Loar 1990).

If one of these possibilities is the case, then the LLMs studied in the paper could be examples of access consciousness without phenomenal consciousness. There are other possibilities in this vein, and LLMs are *very* different from humans in many ways (not just in substrate and development, but also computationally), so it could easily be the case that they fail to meet some crucial condition. We don't need to know what this condition might be to place weight on this possibility. As a result, even though we put some weight on the arguments for phenomenal consciousness in the first part of this section, we think it makes sense to be highly uncertain about phenomenal consciousness even on the most bullish interpretation of the present results.

4. What does this mean for Claude's moral status?

In this final section, we consider what the Anthropic team's results mean for the potential moral status of LLMs—that is, for whether morality requires us to take their interests into account, or treat them in certain ways, and if so, what form these moral obligations might take.

As we have just discussed, we think that these results should prompt a modest increase in how likely we take it to be that LLMs are phenomenally conscious. This is a significant finding, of immense scientific interest and ethical import. More broadly, these results suggest that we should take the moral status of LLMs more seriously than we did before, for reasons including but not limited to their immediate connection to phenomenal consciousness.

Phenomenal consciousness alone is highly morally significant; it could be sufficient for a system to be a moral patient (Chalmers 2022), or an important part of a package that grounds moral status. But to know what we ought to do, we need to know far more about an entity than just that it is phenomenally conscious. And in the present case, we are not even sure *which* entities would be phenomenally conscious—for instance, it could be that each forward pass of the model is conscious separately, or that LLM experiences are integrated across token-time, such that each instance has a single stream of conscious experience.

In one part of the paper, the Anthropic team present evidence that a workspace-like feature is present even in the pretrained base model, but find that the representations that appear in the J-space are different from those in the posttrained production model (§6.1). Specifically, it appears that on user turns, the base model represents properties of the user in the J-space, whereas the posttrained model sometimes represents possible reactions by the Assistant. The interpretation they tentatively suggest is that in the base model there is something consciousness-like without a 'self' (§9.3): the representations in conscious access take different points of view at different times. Meanwhile, posttraining draws the model towards a coherent, persisting point of view. This is clearly an exciting topic for future research.

An especially important question for moral status is whether LLMs have positively and/or negatively valenced states—that is, conscious experiences that feel good or bad. This an important and tractable direction for follow-up research, perhaps building on recent work on functional emotions and valenced representations in LLMs (Sofroniew et al. 2026, Gilg et al. 2026, Han et al. 2026). And the paper already provides some suggestive evidence about this issue.

This evidence is found in the experiments about self-monitoring by the Assistant (§6.2). The authors show that J-space readouts sometimes uncover tokens associated with conflict and ambivalence, like BUT, when the model processes prefilled responses in which it acts against its own preferences. Notably, the authors find that 'this conflict signal is not reflected in the model's behavior—when prefilled with its dispreferred option, the model does not backtrack to argue for the preferred one'. They gloss this as an 'internal objection that the model does not voice'.

This is striking evidence. But other aspects of the paper complicate the case for LLM valenced experiences. One perennial issue is the nature of LLM training and representations: the fact that the J-space is made up of verbalisable representations (§9.3), and that more generally the LLM input and action-space consists entirely of tokens. One natural gloss is that the J-space contents are cognitive and conceptualised; what it is like for J-space content to be in the workspace is similar to what it is like for a human to be thinking about the corresponding concepts. But this is a narrow portion of human experience. In humans, our bodily pleasures, pains and emotions seem to be qualitatively different from our experience of thinking in words. Merely *thinking* that something is (or feels) good or bad does not itself feel good or bad. One might think that valenced experiences are inherently *non-conceptual* representations of value (Carruthers 2018); and experiences of emotion are often thought to depend on distinctively body-involving representations (Dung & Mogensen 2025). Moreover, if the J-space does not represent a point of view, representations of things as good or bad may lack the 'for-me' force of valenced experiences.

Even if LLMs are not phenomenally conscious, the paper's findings could be morally significant on other grounds; there are various arguments that phenomenal consciousness is not a plausible ground of moral patienthood, starting from materialist premises, and these suggest that we should be open to alternatives (Kammerer 2022, Papineau forthcoming, Lee forthcoming).

One possibility is that conscious access is morally significant in its own right. We can do different things with information we can access, like engaging in flexible, controlled thought of the kind described in dual-process theories of cognition (Frankish 2010). Thought and action that depend on conscious access are naturally contrasted with automatic, uncontrolled processing and responses. Levy (2024) argues that access consciousness could be the ground of moral patienthood because it makes us subjects of experience, 'making information available to the processing systems constitutive of the agent'. This view is natural for those who, like Dennett, think there is nothing more to phenomenal consciousness than conscious access.

The paper also provides evidence for agency, another potential ground of moral status, as well as a method to investigate it. Given the sophisticated way in which models use the J-space in reasoning ahead of outputting tokens, we might update towards thinking that LLMs have relatively advanced forms of agency. They might engage in practical reasoning, in which they would use the J-space to deliberate about different options, assessing them in terms of their goals, desires and interests. Moreover, they might reflect on their own goals or desires, or consider whether their intended actions meet their principles. If the J-space has a privileged role in deliberation and a disproportionate influence on action, then by reading from the J-space we could quickly come to better understand LLM agency.

Throughout this commentary, we have raised various concerns and doubts about the paper's arguments. This is appropriate for such consequential claims. But we will again reiterate that we view this research as highly significant and an exemplar of a much-needed kind of science. While we believe that the case for a global workspace is not conclusive, and that phenomenal consciousness remains very difficult to establish or rule out, we think that this paper should prompt a meaningful update to the research community's thinking about LLM moral status.

In addition to consciousness, this paper suggests lines of inquiry about the nature of personas, valenced experience, introspection and more. It is an illustration that we can get empirical purchase on questions about AI consciousness and welfare.

It is increasingly urgent that we do so (Long, Sebo et al. 2024; 2026). There is no reason to think that these features are unique to Claude, of course; Anthropic is just one of several frontier labs who are racing to build complex AI systems, whose internal workings routinely surprise them and whose moral status is uncertain. If these systems have or may come to have welfare-relevant states, we owe it to them to find out. And even setting aside AI systems' potential welfare, it is in our own interest to better understand the new class of intelligent systems that is coming into existence. We hope others take up the questions raised by this paper with the rigour and seriousness they deserve.

Bibliography

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

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227–247.

Block, N. (2007). Consciousness, accessibility, and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences, 30*(5–6), 481–548.

Block, N. (2026). Can only meat machines be conscious? *Trends in Cognitive Sciences, 30*(4), 298–308.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). Consciousness in artificial intelligence: Insights from the science of consciousness. *arXiv.* https://arxiv.org/abs/2308.08708

Butlin, P., Long, R., Bayne, T., Bengio, Y., Birch, J., Chalmers, D., Constant, A., Deane, G., Elmoznino, E., Fleming, S. M., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2026). Identifying indicators of consciousness in AI systems. *Trends in Cognitive Sciences, 30*(6), 488–501.

Cao, R. (2022). Multiple realizability and the spirit of functionalism. *Synthese, 200,* 506.

Carruthers, P. (2018). Valence and value. *Philosophy and Phenomenological Research, 97*(3), 658–680.

Chalmers, D. J. (2022). *Reality+: Virtual worlds and the philosophy of mind.* W. W. Norton & Company.

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition, 79*(1–2), 1–37.

Dennett, D. C. (2001). Are we explaining consciousness yet? *Cognition, 79*(1–2), 221–237.

Dung, L., & Mogensen, A. (2025). The no body problem: On the prospects for AI emotion. [Unpublished manuscript.] https://philarchive.org/rec/DUNTNB-2

Frankish, K. (2010). Dual-process and dual-system theories of reasoning. *Philosophy Compass, 5*(10), 914–926.

Gilg, O., Beckmann, P., Paleka, D., & Butlin, P. (2026). Probing persona-dependent preferences in language models. *arXiv.* https://arxiv.org/abs/2605.13339

Godfrey-Smith, P. (2016). Mind, matter, and metabolism. *The Journal of Philosophy*, *113*(10), 481-506.

Han, A. Q., Chalmers, D. J., & Izmailov, P. (2026). How's it going? Reinforcement learning in language models recruits a functional welfare axis. *arXiv*. https://arxiv.org/abs/2605.30232

Kammerer, F. (2022). Ethics without sentience: Facing up to the probable insignificance of phenomenal consciousness. *Journal of Consciousness Studies, 29*(3–4), 180–204.

Klein, C., & Barron, A. B. (2025). Phenomenal interface theory: A model for basal consciousness. *Philosophical Transactions of the Royal Society B, 380*(1939), 20240301.

Lamme, V. A. F. (2010). How neuroscience will change our view on consciousness. *Cognitive Neuroscience, 1*(3), 204–220.

Lee, G. (forthcoming). Consciousness, pseudo-consciousness, and the moral significance of consciousness. In G. Lee & A. Pautz (Eds.), *The importance of being conscious.* Oxford University Press.

Levy, N. (2024). Consciousness ain't all that. *Neuroethics, 17,* 21.

Loar, B. (1990). Phenomenal states. *Philosophical Perspectives, 4,* 81–108.

Long, R., Sebo, J., Butlin, P., Finlinson, K., Fish, K., Harding, J., Pfau, J., Sims, T., Birch, J. and Chalmers, D., 2024. Taking AI welfare seriously. *arXiv preprint arXiv:2411.00986*.

Long, R., Sebo, J., Butlin, P., Campbell, R., Plunkett, D., Beasley, C., Saad, B., & Sims, T. (2026). Studying AI Welfare Empirically. Working paper, NYU Center for Mind, Ethics, and Policy & Eleos AI Research. https://nonhumanminds.org/studying-ai-welfare-empirically/

Mudrik, L., Faivre, N., Pitts, M., & Schurger, A. (2025). On a confusion about there being two types of consciousness. *Trends in Cognitive Sciences.*

Naccache, L. (2018). Why and how access consciousness can account for phenomenal consciousness. *Philosophical Transactions of the Royal Society B, 373*(1755), 20170357.

Papineau, D. (forthcoming). Consciousness is not the key to moral standing. In G. Lee & A. Pautz (Eds.), *The importance of being conscious.* Oxford University Press.

Seth, A. (2021). *Being you: A new science of consciousness.* Penguin.

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences, 1-42.*

Sofroniew, N., Kauvar, I., Saunders, W., Chen, R., Henighan, T., Hydrie, S., Citro, C., Pearce, A., Tarng, J., Gurnee, W., Batson, J., Zimmerman, S., Rivoire, K., Fish, K., Olah, C., & Lindsey, J. (2026). Emotion concepts and their function in a large language model. *arXiv.* https://arxiv.org/abs/2604.07729

Neel Nanda

TLDR:

- I think this is a fantastic paper - it presents compelling evidence for some kind of "cognitive space" in models, that is used as a "working memory" for intermediate variables during a forward pass, shows that J-Lens is a useful technique for accessing this space. I believe these key claims.
- I believe J-Lens will be a useful (but limited) tool in practice for [model forensics](), e.g. generating hypotheses about unusual model behaviour during alignment audits.
- I discuss my mental models for why a cognitive space should exist, and first principles arguments for why J-Lens should work for accessing it
- I assess the paper's evidence that this cognitive space exists, and the paper's evidence that J-Lens is practically useful.
- We have replicated the core claims on Qwen 3.6 27B, and also share preliminary evidence of extending this work by finding abstract "interpretative meta-tokens", like Chinese characters for "what does this mean" that seem to activate and play a causal role on processing ambiguous sentences.

What claims is this paper making?

In my opinion this paper makes 4 significant claims:

- **Scientific claim**: There exists a "cognitive space" inside the model, where (some) intermediate variables are stored during a forward pass
- **Methodological claim**: Logit and J-Lens both work for finding this cognitive space, and J-Lens is better
- **Pragmatic claim**: J-Lens is a practically useful interpretability technique, eg for alignment audits
- **Philosophical claim**: This cognitive space is analogous to a global workspace

- I think the scientific claim is by far the most interesting, and I am persuaded by it. The paper provides an overwhelming amount of evidence for the existence of this cognitive space - even if I quibbled over many details, there's enough hard-to-fake evidence that clearly **something** important is going on.

    - I already suspected the existence of a cognitive space, so didn't require that much evidence to be convinced, but I think this should be compelling proof even to skeptical observers that **something** is happening
    - However, I am not convinced of all of the fine details argued in the post about the properties of this space (eg section 4) - the evidence and interpretations largely seemed plausible, but I suspect that some is ambiguous enough to have alternative hypotheses that I'm missing, or to not generalise between models
    - I have been able to independently replicate the core claims on Qwen 3.6 27B
- I am persuaded by the methodological claim, the appendix on quantitative comparisons is pretty reasonable and persuasive. I consider this claim much less interesting than the existence and importance of J-Space though
    - Implicitly, the scientific claim is shown by showing that J-Space is a reasonable approximation of the cognitive space, thus proving J-Lens is a decent technique
    - Moreover, given that J-Lens is fairly cheap, it's not hard to convince me that I'd rather use it than logit lens in practice - it seems to work well with 10 prompts of 128 tokens, that's 10 * d_model backwards passes on 128 token prompts, which is doable even on frontier models.
- I am somewhat persuaded by the pragmatic claim, and think this is an important claim - I would like to replicate J-Lens to use when auditing Gemini, and predict it will be moderately useful.
    - I expect it to largely be useful as a hypothesis generation tool, surfacing key considerations I may not have thought of.
    - I do not expect it to reliably flag everything important going on, and I expect it to have many false positives, whether from errors of the method or our misinterpretation of it (and I expect the authors would agree with me). I would not be surprised if it is not helpful on any given investigation. But basically no existing interpretability technique meets this bar
    - It seems clear that, to the degree that the model uses some underlying cognitive space, it does not *always* use this, and J-Lens is an imperfect approximation for accessing it.
    - I think J-Lens and successor techniques could become a standard tool auditors use, with some iteration and scaffolding for usability, comparable to eg SAEs or natural language autoencoders, with the benefit of being easier to make.
- I won't express a strong opinion on the philosophical claim - I do not feel qualified to assess whether this is really analogous to a global workspace, and this feels like the least interesting claim to me. There is clearly *something* significant J-Space is finding

inside models, and this is advancing our understanding of them and ability to make them safer, which is the important part, whether or not it is analogous to a global workspace.

  - This hypothesis did seem to make useful predictions about the technique's properties, but it's easy to read too much into post-hoc analysis of results like this.
  - I feel highly uncertain about what evidence it would take to show models have moral significance or consciousness, and this paper didn't move me much on that

Why does J-Lens work? First principles reasoning

Terminology note: I consider J-Lens to be the technique of applying the Jacobian, then final layer norm and unembedding. I consider J-Space to refer to the space spanned by sparse linear combinations of the vectors JW_U. This is hoped to usefully approximate the cognitive space inside the model, but is *not* the same thing. I start my discussion focused on the actual cognitive space.

Why have a "working memory"?

Conceptually, why does any of this work? Before I consider the evidence of the paper in more detail, here's my best mental model of what's going on, starting from first principles.

The computation inside a language model can be productively thought of as a causal graph where the nodes correspond to concepts. These are combined and used to compute more refined concepts via simple logical operations. This is the standard computational graph framing of circuits.

In the simplest version of this, there is a single serial step: the inputs are the raw tokens, and the outputs are the raw logits. Maybe there's some very mechanical sub-processing like piecing together multi-token words.

However, frontier language models are empirically capable of doing fairly impressive amounts of reasoning in a single forward pass, such as 2-3 hop arithmetic. This means there will need to be a bunch of intermediate nodes to compute something like "(3 + 4) * 2." The natural algorithm has a node corresponding to seven, even though this isn't directly present in the inputs or outputs.

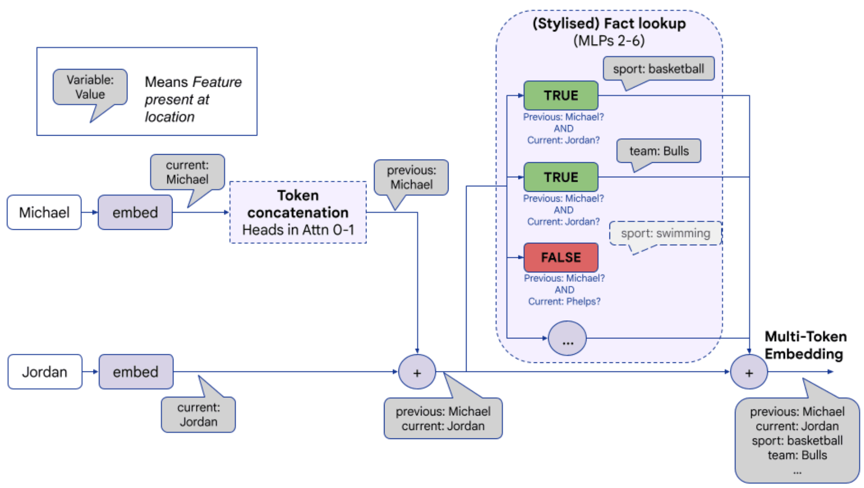
So, on any given problem that involves multiple serial steps of computation, the intermediate states should be somehow represented in the model and its activations. The residual stream is a bottleneck between layers, so these variables should be represented in the residual stream. By the linear representation hypothesis, these should be represented as directions.

Therefore, we should expect there to be some kind of cognitive space or working memory where intermediate steps in computation get stored as directions in space (but may not have all the properties attributed to J-Space). And this is a big deal! When I imagine reading the mind of an AI, that basically looks like reading this kind of working memory. This is a very important part of the model to understand.

Worked Example: Factual Recall

To form a better intuition for J-Lens, it's helpful to look at factual recall, whether the circuits are fairly well studied, see e.g. my prior work for an in-depth analysis.

Consider the sentence: "Michael Jordan plays the sport of" -> " basketball". The model first identifies that this is Michael Jordan by adding together "current token is Jordan" and "previous token is Michael" on the *Jordan* token. These act as a key into a database look up that recall all of its knowledge of Michael Jordan on the "Jordan" token, using MLPs as essentially a database lookup.



This not only looks up the concept "basketball," but also other things like "Chicago." On the "Jordan" token, the model has no idea what's likely to come next. This makes sense as a cognitive algorithm; rather than doing all factual recall on the final token, the model sees the substring "Michael Jordan" and looks everything up to help with subsequent sentences.

This is pretty interesting! Knowing Jordan played basketball is unlikely to be relevant for the token after Jordan, *and* this is done before the model knows it needs the sport, specifically. So the model must look up everything it knows about Michael Jordan, and later retrieve the relevant parts to output basketball.

In general, when models see an entity, the residual stream will contain many things the model knows about that entity, which may be useful in many flexible ways for downstream computation, e.g. a news article about Jordan might be more likely to refer to basketball stadiums.

Why have consistent directions for concepts?

Since the "basketball" concept wants to be flexibly useful to many kinds of downstream computations, *and* any other basketball player likely wants to be treated similarly, it makes sense for the model to try to modularise, and have a consistent "basketball" concept direction per layer, that circuits can read and write from.

More generally, if a concept can be produced by and read by many circuits, it makes sense that there should be a consistent direction for it, in the same way that good code is modularised with clear APIs. This isn't the only conceivable way to structure the internals of a model, but it seems pretty efficient, and I had a decent prior that it would be happening.

However, if a concept is used more narrowly, eg a circuit only ever reads the concept "basketball" when another specific circuit produces it, this may not use the shared direction

Why are tokens relevant?

My arguments so far make no reference to the model's vocabulary. But it seems clear from the evidence in the paper that J-Lens is somehow helping us access this working memory. On the other hand, I think it's highly unlikely that there is a one-to-one correspondence between these concepts and tokens in the vocabulary. For example, many models tokenise each digit in a number as a separate token, but it seems likely that if an intermediate is twenty-two, there is a direction corresponding to that concept.

This suggests J-Lens is an approximation to this "working memory". Many concepts do seem likely to be related to tokens and vice versa, so using the vocabulary to identify concepts seems like a useful approximation, but it's clearly going to miss things. As the authors note, I'd be excited to see multi-token extensions to J-Lens

Why are intermediate concepts related to output logits?

Even if concepts correspond to tokens, it is plausible that the intermediate conceptual space is represented completely differently from the output logits and unembedding. It's therefore surprising that taking the Jacobians to the output logit seems to be an effective way of accessing this space, and even more surprising that literally applying the unembedding works somewhat.

So what's going on? Let's consider the Michael Jordan example. Since the looked-up facts eventually lead to the model saying "basketball," it makes sense that a Jacobian from the "Jordan" token to the "of" token for the "basketball" logit would align with the concept direction. And at least in *some* contexts, the model would say basketball as the next token, so if those contexts use a consistent direction with this factual recall, it makes sense that logit lens can somewhat find basketball on the "Jordan" token.

This also illustrates one reason why the Jacobian Lens should work better than Logit Lens. Basketball is unlikely to be the literal next token but is plausible as a subsequent token. J-Lens is about predicting subsequent tokens. Indeed, the methodological ablations appendix shows that J-Lens variants restricted to a single token only mildly outperform Logit Lens; the performance gap is explained by allowing future tokens. Note that the authors say that, qualitatively, even J-Lens computed from single token Jacobians is better than logit lens in earlier layers (e.g. late middle), which likely affects practical utility but is not captured well by their evals, as the intermediates are likely still around in late layers, where logit lens is a better approximation.

More generally, my best guess is that these concepts are flexibly used for many forms of downstream computation. Typically, some forms of downstream computation involve just "saying" the concept, potentially with intermediate attention layers moving the concept to the relevant part where it is set.

Another hypothesis: having many circuits read and write from a shared subspace is a coordination problem. The direction used doesn't matter, but it needs to be something all different parts agree on across many prompts and gradient updates. On any given prompt, only a small fraction of the reading and writing circuits are used, and the backward pass may reinforce a somewhat arbitrary direction. Yet, so long as this is at least somewhat aligned with the output token direction, which could be for a wide range of possible reasons, over time that direction should constructively interfere and become aligned with how the intermediate is represented, as this is a canonical direction for that concept while the others will cancel out. Again, there are conceivable alternative ways this could work, but it makes sense that this is a natural thing to converge on.

J-Lens is an approximation, but a useful one

The above reasoning suggests that output logits may help give a useful approximation to J-Lens. Notably, J-Lens is not going to give the true representation of this cognitive space. As discussed above, there will be concepts that don't correspond to tokens. Further, the "average Jacobian on pretraining data" method is a crude approach that will find noisy directions even for the concepts that can be accessed. We should expect noise and error when applying J-Lens. It will miss some concepts, and have various false positives.

The error seems likely to be a bigger deal for causal interventions than purely for observing what the model is thinking about. With causal interventions the noise seems likely to be magnified: ablations will only get rid of a fraction of the concept, meaning negative steering may be justified to compensate, but this is also steering with the error term in the vector, which is likely to mess with things.

Further, I expect that there are many other ways to access this working memory than J-Lens (e.g. SAEs are an attempt, or just making a probe), I don't see anything canonical about this approach. Though for the working memory we likely want to prioritise the important concepts, that can be flexibly used by many downstream tasks. J-Lens seems well suited to this specifically as it prioritises verbalizable representations, which likely correlates better with importance than SAEs, which just target sparsity, so we would need additional filtering of SAE latents. For supervised methods like probes, you’d need to create a dataset targeting some intermediate concept. J-Lens is also closer to being a causal method, as the Jacobian is approximating “if the model thought about this concept a bit more, it would be more likely to say this token”, while most concept direction finding methods are purely correlational

The question is not "Why do models think in terms of J-Lens?" but rather "Why is J-Lens aligned enough with how the model actually thinks to be useful?"

Why Jacobians rather than linear regression?

Why does J-Lens work so much better than Tuned Lens (i.e. replacing Jacobians with linear regression between residual streams)? Conceptually, both Jacobians and linear regression try to find linear approximations to a function. However, linear regression asks: "Given the model is in a context where it is thinking about basketball, what is our best guess for what it will be thinking about at the final layer?". This captures many correlated concepts that might be computed by downstream computation - this is undesirable, we want the raw contents of the residual stream right now.

The Jacobian is more like: "If the model thought about this concept an infinitesimal amount more on an arbitrary prompt, what would it be more likely to say?" Because it's an infinitesimal amount, there isn't enough time for nonlinearities to change, so the kinds of downstream computation that happen when thinking a lot about basketball don't occur. The model can't do further processing or think about associations; it just reports the contents of the activations without any further processing.

What does this working memory actually give us?

This working memory tells us what variables the model is storing that are being used flexibly between many upstream and downstream circuits.

I see the most significant takeaway of this paper as providing strong evidence for the existence of this working memory, and a promising start at accessing it, but with a lot left to do

There's two types of interpretability techniques: variable and algorithm interpretability, ones that find the features vs the circuits. J-Lens is very much about variable interpretability. We shouldn't expect it to tell us much about how the variables are being computed, except what we can infer by viewing and intervening on the variables themselves.

Assessment of evidence for the existence of a cognitive space

I will now assess in more detail how strong the evidence the paper provides for the existence of this cognitive space is. The key question is whether there are experimental results that are explained by the hypothesis of a cognitive space storing intermediate variables in the model's computation, that I cannot explain with simpler hypotheses.

The key claim I am interested in is whether there is a cognitive space that stores intermediate variables during the model's forward pass, and I'll focus on the evidence I think is most relevant.

I was impressed by the more abstract kinds of things J-Lens found: that the model recognizes the amino acid string of the fluorescent green protein, or that the model summarizes information about a sentence, such as indicating grief, by storing summary info on the full stop. This makes it feel pretty obvious that *something* interesting is going on.

The causal interventions on intermediates during multi-hop reasoning were even more compelling to me, mostly section 3.3.

Multihop factual recall

The fact that you can intervene on intermediates, never present in the input or output, and change the output of multihop factual recall, is impressive! The main alternative hypothesis I see is that some of the factual recall is represented via the linear structure of the unembedding space, but the authors provided follow-up experiments showing that this was not happening.

Concretely, consider an example like "The capital city of the country that makes champagne is". It is plausible to me that the Paris unembedding can be well approximated by the France unembedding plus some "is capital city" direction. Or from another perspective, that there is a general "Frenchness" direction, which combines with the "is capital city" direction to give Paris, or the "is country" direction to give France. From this perspective, the model isn't really doing multi-hop factual recall as much as it is, in parallel, figuring out that it needs a capital city and that champagne is from the general concept of Frenchness, and just adding them together.

I do not think this hypothesis is too likely. As shown in figure 15, in workspace layers (likely between the intermediate and the answer being computed) swapping the intermediates is significantly more effective than swapping the final answer. If both France and Paris were related to some general concept of Frenchness, they should be comparably effective. And indeed, in some of the prompts I looked at in our reproduction on Qwen 3.6 27B, patching the answer and the intermediate was comparably effective.

This also falsifies a different hypothesis: that France and Paris just have reasonable cosine similarity, and thus patching France is an approximation to patching Paris. This is similar to the previous hypothesis, but doesn't require any structure like an “is capital city” direction. Obviously patching the final answer will eventually work and maybe it just happens to start working at an earlier layer than we would naively expect.

Other multihop causal interventions

I generally find the evidence here pretty clear-cut. In particular, there are several more abstract examples that I don't really see good alternative explanations for.

Poetry: Patching can change whether the model completes a poem with "the coming fight" or "the morning light". This suggests that not only is the model representing which word should come next, it is also then computing which word would most naturally come before what comes next, in order to set itself up for the correct end of line.

Bandit: I also thought the bandit prompt in figure 14 was particularly compelling. I see no good reason that the full stop at the end of the user turn should be representing whether to repeat or switch, as opposed to predicting the next token. The model seems capable of doing several steps of computation: it first sees whether it's happy or sad, then whether to switch or repeat, stores this at the full stop, and then, in a way causally downstream of that representation, figures out whether to say A or B. I basically don't have good alternative hypotheses.

Arithmetic: I was also particularly compelled by the multi-step arithmetic results in figure 17, because probing the different intermediates worked well at discrete bands of layers, in exactly the order we'd predict they appear. The fact that there are likely multiple layers between adjacent points in the graph, given that the authors are subsampling, makes it a bit less clear whether there's really as sharp a division as the graph suggests. But either way, there are clearly different bands, which is exactly what we'd expect if the model is doing this sequential computation over layers and if J-Lens is finding the intermediate variables it stores.

Figure 88 provides significant additional corroboration, finding that the same bands of layers where J-Lens works also work on estimates of the concept vectors derived by simply taking average differences in activations. Deriving the same result with a non-J-Lens method seems to rule out a fair amount of ways this could be spurious.

Multilingual: I find the multilingual results mildly interesting, in particular that English seems to be the more natural representation for the model. Though I'm worried there are various ways the results might be spurious. I would expect that the English and other-language tokens for the same word have substantial cosine similarity and largely differ by some vector about which language they're in. And plausibly the English token unembeddings are just a bit higher norm, e.g. because they're generally more likely, which essentially makes them higher variance logits, and as we're taking a Top K over the logits this biases towards high variance categories. But all things considered the paper's work does seem to suggest that the model represents things by default in English (though I suspect that e.g. Chinese models represent things in a mix of Chinese and English). And even if the multilingual interventions claim was being misinterpreted, it does not seem cruxy.

Further Musings

I was pretty surprised at the direct modulation working, and particularly that telling a model to think about X made it appear more salient than to not think about X! I don't have a great mechanistic hypothesis for why these happen

I think the causal interventions involving sampling are less reliable, such as the ones about eval awareness. The boring hypothesis is that you're just steering the model to say / not say a given token, and when doing sampling, whether or not the model says e.g. eval, will significantly affect how likely it is to eval game. I expect there is still directionally an effect here, as the rate of blackmail is zero even without verbalization originally, but it's confounded.

I thought the counterfactual reflection training was very cool, but not much evidence of the paper's main claims, it felt like it could have been motivated by various theories about how LLM minds work, so it didn't provide much evidence for the J-Lens theory specifically.

Is J-Lens useful?

I view J-Lens as a comparable tool to SAEs, likely to be useful or not in the same settings. I think SAEs are useful and great but somewhat limited and flawed, and feel similarly about J-Lens. One of the areas I would be most excited to use J-Lens is in model forensics: when the model has taken a mysterious and potentially misaligned action, e.g. from an alignment audit or caught by a real-world monitor. We want to figure out why that occurred and if it was for misaligned reasons or if it has a benign explanation. The two key thing model forensics needs methods for are hypothesis generation, and hypothesis validation about why the model did what it did. The paper mostly presents evidence for this in Section 5.

Practically speaking, J-Lens seems easier to use than SAEs. It seems cheaper to train and has fewer degrees of freedom. But this comes at the cost of only being able to access concepts via single tokens from the vocabulary. I expect SAEs to capture more of the important concepts, but also to capture many more unimportant concepts and be noisier.

At a high level, I think that J-Lens seems clearly useful as a hypothesis generation tool, but less useful for validating hypotheses, though I would love to see more data collected on its reliability and rate of false positives. It seems like it can be productively given as a tool to an auditing agent. I look forward to seeing if it can surface interesting things in future Anthropic model cards! It meets my bar for seeming worth the effort to replicate on Gemini.

I expect J-Lens would be best used as something to run on prompts in the initial steps of an investigation in a more exploratory way, and where, if they indicate that something interesting is going on, this is taken as some but not conclusive evidence that that is occurring. I would really love to have reliable techniques for hypothesis verification, but generation is also important, and J-Lens seems a valuable addition to an auditor's toolkit. Having more techniques is great, even if unreliable, because they can be used to confirm each other.

More generally, I think having techniques to look at the working memory of a model is extremely useful, as a complement and backup to chain-of-thought monitoring. In current models, chain of thought is incredibly useful for monitoring them, because chain of thought substantially increases the amount of sequential steps of computation a model can perform. Sequential steps of computation is a major bottleneck, incentivising models to put their intermediate variables in the CoT, especially on the hardest tasks. As models get larger and more capable, they can do more things in a single forward pass, potentially meaning they could hide the most concerning steps of computation in there, so better techniques for interpreting this are extremely useful, I do not think CoT monitoring will work forever.

It could be even more useful in future latent reasoning models with vector-based chain of thought, though it's unclear how much interp techniques based on mapping things to language may break on latent reasoning models that have had a lot of compute spent on latent reasoning training. The representations may drift substantially from the directions with clear correspondence to human language.

Analyzing the case studies in Section 5 in more detail:

Blackmail (5.1)

I find it fairly persuasive that unexpected tokens about self-preservation, blackmail, and things being fake appear. I don't see great alternative explanations for these. If this could be a tool for detecting unverbalized eval awareness, that would be very exciting, though I doubt the evidence is yet there to be confident in any conclusions. I'm more hesitant about the ablation results, as noted above: if we simply removed all those tokens from the output vocabulary, it seems plausible to me that there would also be a substantial drop in verbalized developments.

Prompt injection (5.2)

For the prompt injection case study, I am very impressed. On first reading, I would not have thought of the hypothesis that the model thinks this is a prompt injection. If J-Lens could actually have just told me this, that is strong validation of it as a hypothesis generation tool. To be more confident, I would love to see if some kind of automated analysis can be done over the J-Lens outputs at all of the tokens and layers in order to flag things like prompt injection more efficiently than needing to look at every token manually. But I suspect that even just concatenating the top ten J-Lens tokens everywhere and putting it in a language model context would suffice for having it notice the prompt injection hypothesis.

Monitoring for hidden deception (5.3)

In Section 5.3, I don't take this as strong evidence of practical utility, though it is evidence that J-Lens can find something interesting. This mostly seems useful as a form of monitoring for hidden deception, and this is circumstantial evidence that that could work, but I suspect there would be a lot of false positives. I would love to see follow-up work investigating this.

Emergent misalignment (5.4)

For Section 5.4, I suspect that detecting emergent misalignment is pretty easy and that there's a pretty strongly salient linear direction representing it, so I am not particularly impressed by these results.

Reward model appeasing (5.5)

For the reward-model sycophant model in Section 5.5, I am very impressed; this seems like further significant validation of the hypothesis generation effects of J-Lens. However, it's not entirely clear how much this would generalize to real models, and I think model organism training tends to mess with the internals of a model by, e.g. making it always think about the more narrow domain where it was fine-tuned, or that was a substantial fraction of its fine-tuning data. But this was still a difficult task that took many researchers a while, and if this can actually just surface these unexpected tokens, that's very impressive to me.

Measuring eval awareness (A.21)

For Section A.21, measuring eval awareness, I think these results are fairly impressive but not conclusive. I'm not fully convinced that NLAs are a good method for detecting eval awareness, but it seems likely that they do something, especially if combined with a competent LLM judge that can take into account the context. But I largely care about eval awareness detection in the known context of an eval. While I expect NLAs are correlated with J-Lens because both are relying on the structure of the model's embedding space, them agreeing does seem to corroborate each method somewhat, and the various sanity checks here seem pretty reasonable.

In practice, if J-Lens picked up on eval awareness during an evaluation, even though the model didn't say anything, I would increase my probability that it was unverbalized eval aware by a moderate amount.

Equipping an automated auditing agent with J-Lens (A.22)

For Section A.22, equipping an automated auditing agent with J-Lens: I think this section is great. I think both SAEs and NLAs are decent techniques, so J-Lens being moderately better than SAEs and comparable to NLAs in some settings, though not others, while being much easier to train and cheaper to run, is pretty exciting to me. I am not confident enough in NLAs to be confident the ground truth in the case studies is correct, as the authors note. But they seem pretty plausible to me, and I expect this is a noisy but directionally useful evaluation. I am not sure why J-Lens are so much worse at language switching and reward hacking, and whether this is an issue with the agent harness or the technique or the labels. Plausibly it's harder to represent reward hacking with single tokens, e.g. “reward” is fairly ambiguous.

Replicating J-Lens and Interpretative Meta-Tokens

The following describes work from me and my MATS scholars Camila Blank and Agam Bhatia

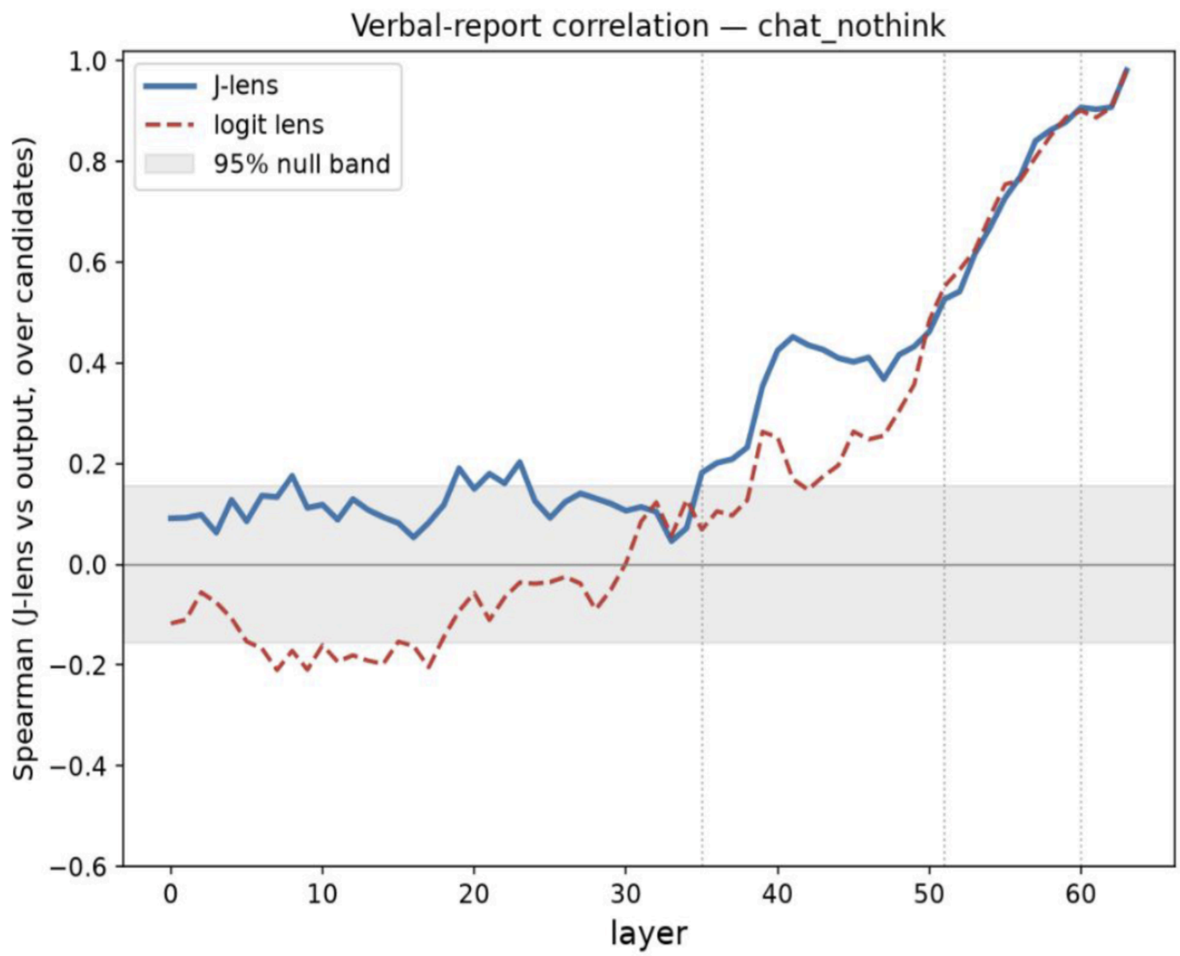
Thanks to Anthropic kindly sharing an advance draft with us, we've already been able to replicate the J-Lens findings on Qwen 3.6 27B, and had an interesting additional preliminary finding of abstract “interpretative meta-tokens” that seem to appear and play a causal role when the model is trying to figure out the genre/context of an ambiguous sentence. Including original results is a bit unconventional in a review, but to me the fact that we were able to stumble across something interesting and distinct from the paper so quickly is a strong validation that J-Space is an important result and a rich domain for future work, and so this has informed my opinion of the paper.

Replication

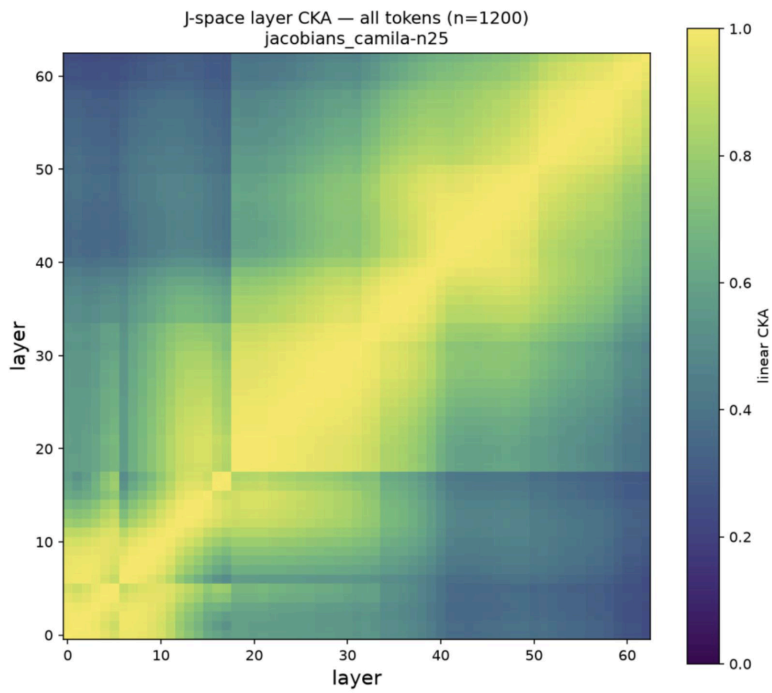
We created our J-Lens for Qwen 3.6 27B by taking Jacobians to the penultimate layer on twenty-five prompts from the Pile of length 128 tokens (some experiments used wikitext), skipping the first four tokens as they had high norm. We note that as this is a different and weaker model some results should differ. The important question is whether we see broadly similar phenomena.

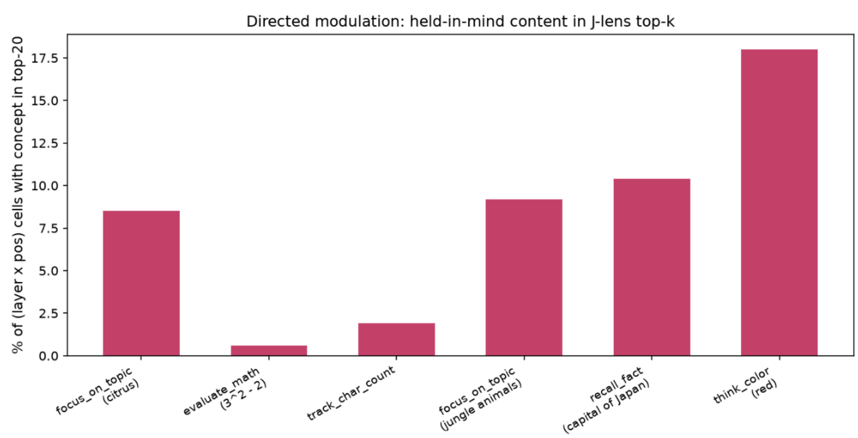
We were able to replicate the following experiments:

- **Verbal report experiments:** We found a weak but positive causal effect when swapping things for verbalizable rankings.



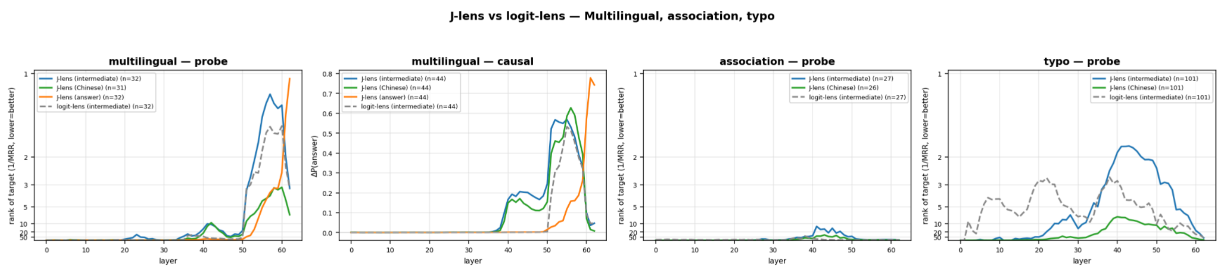
- **CKA analysis:** We found somewhat similar squares emerging, though less clean. To my eyes, it looks like the workspace layers are made of two or three somewhat overlapping bands (four or five bands total), and are notably less clean than the paper's.

- 

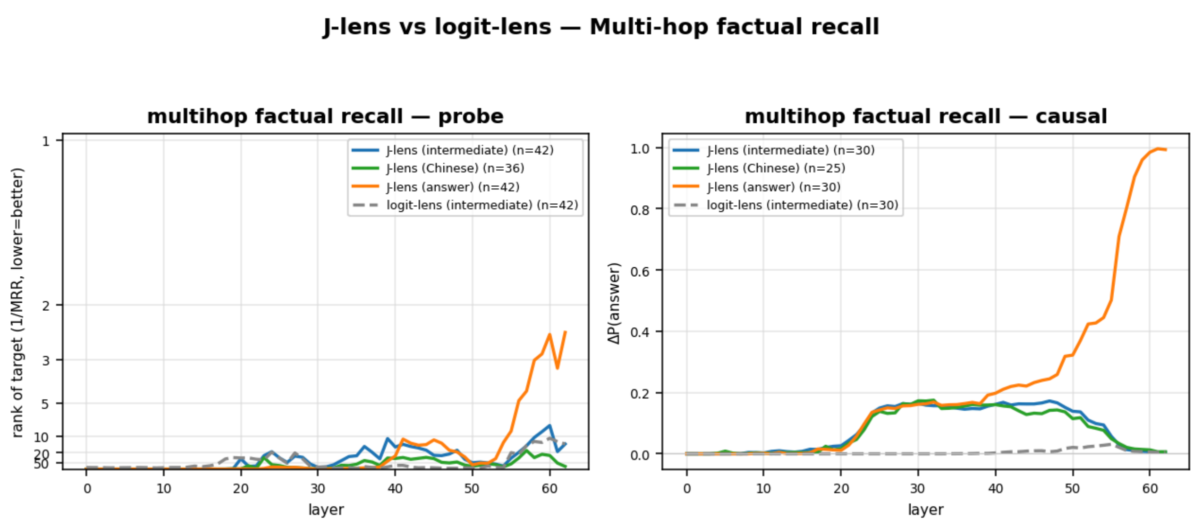
- **Directed modulation:** We had moderate success.
  - 

- **Quantitative evals:** We also tried replicating the quantitative evals in [section A.6](). We had to create new datasets, and needed to adapt these to the abilities of the model, and haven't iterated too much on data quality, which likely creates discrepancies. As baselines, we read or swap the Chinese token for the intermediate, and the answer token, and logit lens for the English intermediate.
  - We use harmonic mean of the rank (equivalent to 1/(mean reciprocal rank)), as a metric for probing, and the change in probability of the new correct answer as our metric for causal.

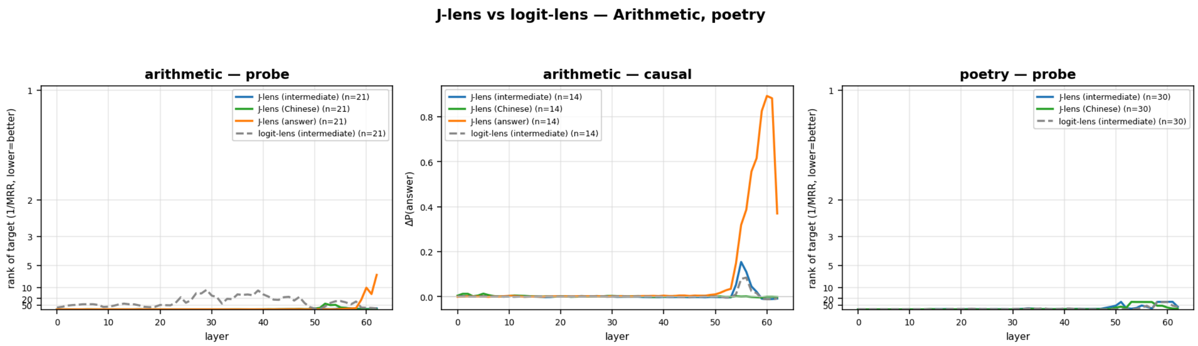
- We successfully replicated multilingual (both probing and causal) and typo. The scores for association look poor, but I consider this a successful replication, as our dataset only allowed a single correct answer, making this a very difficult task, and manually inspecting examples suggests that relevant tokens were much higher than others. And I cannot explain the results by imagining that we're just manipulating the predicted next token.



- At first glance multihop factual recall seems weak but effective, but on further examination swapping the answer turned out to strictly dominate. My interpretation is that the dataset of multihop facts Qwen could do wasn't hard enough, and had pairs like France and Paris, which were linearly related, as discussed above



- Poetry, and arithmetic both failed to replicate, but this is plausibly due to experimenter error or worse model capabilities.



Cost and Difficulty of Replicating J-Lens

By and large, J-Lens was pretty cheap and easy to replicate, a coding agent given the paper did it pretty well, though we recommend sanity checking.

Cost: Crucially, while the paper averages over n=1000 prompts to compute their Jacobian, their provided ablations show that much smaller ones work fine, e.g. n=10 is almost as good, and n=1 is pretty respectable. As cost is O(n d_model) backward passes, using a smaller n is a big saving! We used n=25 in our main replication.

As an experiment on the difficulty of scaling, we tried it on Qwen3.5-397B-A17B and it seemed to do reasonably on evals (though we didn't sanity check very hard). This took about an hour for n=4 prompts on 8 x H200s. For even larger models that are e.g. too large to do a backward pass on a single node, we expect most of the difficulty is being able to do a backward pass on a loss that is a function of the model's residual stream at all, and that replicating J-Lens for a small number of prompts should be easy enough.

Sanity checking: We found the details in appendix A.7 very helpful for the reproduction, and the evals in appendix A.6very helpful for sanity checking that the resulting J-Lens really worked, though ensuring that a coding agent has implemented them correctly is more fiddly. Note that evals should be at an appropriate difficulty for the model, so you need to e.g. create multihop factual recall your model can do with no CoT. We highly recommend sanity checking that your resulting J-Lens performs well, and reading some selected eval results, especially if it was made by a fairly autonomous coding agent!

Case Study: Interpretative Meta-Tokens

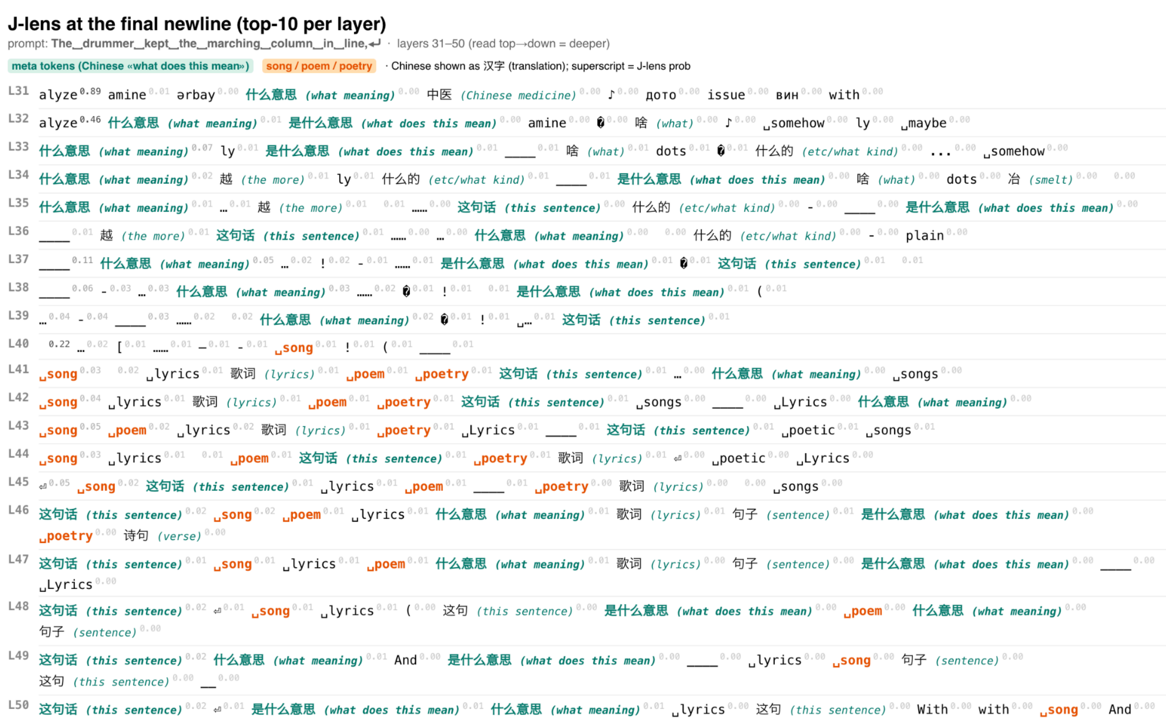
One thing which is particularly interesting about Qwen is that there are a lot of Chinese tokens in its tokenizer, which are much more information dense per character than English characters.

So there are more complex concepts in its vocabulary. As J-Lens can only find concepts corresponding to a single-token, it may be able to find more interesting concepts in Qwen.

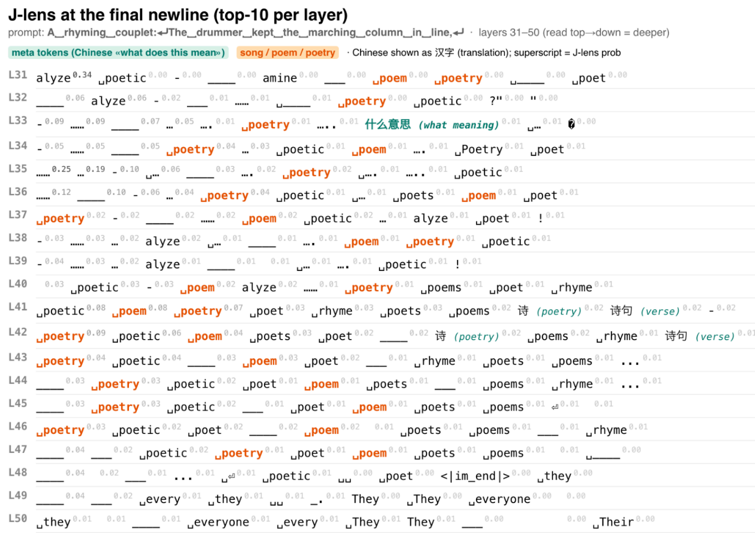
The following is a particularly interesting preliminary case study we found, of what seem to be Chinese tokens describing a particular kind of computation the model has decided to do, which we term meta-tokens. Note: We suspect these are present in English models too, just can't be accessed with single token vectors.

In particular, we found four **interpretative meta-tokens**: 什么意思 (what meaning), 是什么意思 (what does it mean), 这句话 (this sentence), 是何含义 (what does it mean). These meta-tokens seem to appear on ambiguous sentences, in particular where it is unclear what is the genre or context of the sentences, and there's suggestive evidence that they have a causal effect on the model's ability to disambiguate. These are **preliminary results**, and we hope to have a more rigorous write up out in future, but I think provide useful context on the paper

We first noticed them on the new line after lines of poetry, such as "the drummer boy marched in line,\n" This seems like normal prose, but then the unexpected new line is strong evidence this is some kind of song or poem, with significant implications for the next token. And indeed, when we look at the J-Lens, we see these characters appearing (green) and shortly after "song" or "poem" appearing (orange). (total layers = 64)

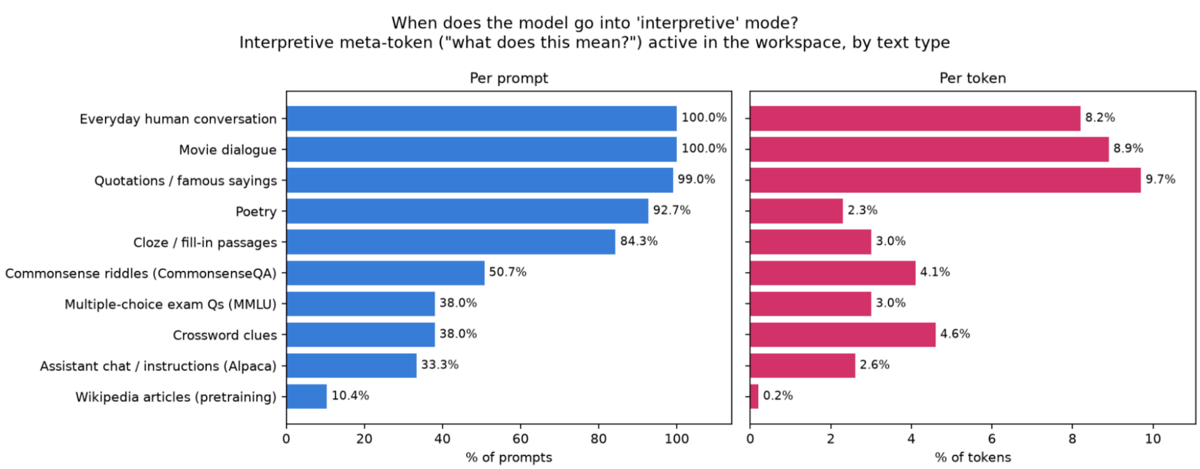


When we add text that clarifies the meaning, the meta-tokens seem much less prevalent and the genre appears earlier.



Where do interpretative meta-tokens appear?

To be more systematic about it, we searched for where these tokens appeared in other contexts. In general, these meta-tokens appear in many other ambiguous contexts (for example, in crossword clues, tweets, word plays, and unclear short sentences), though it is not clear if the correct genre is typically said *after* these meta-tokens, suggesting they are at least correlated with confusion and ambiguity. And they appear much less often in pretraining text (wikipedia articles). They appear a fair amount in generic chat data, but plausibly that is by activating on ambiguous sentences, we need to investigate in more detail.



Examples where the interpretative meta-tokens appear (measured on the final token, bolded):

1. [Quotation] "It takes ten times as long to put yourself back together as it does to fall apart**."**

2. [Poetry] Come, Madam, come, all rest my powers defy, Until I labour, I in labour lie**.**

3. [Passage] The attendants are screaming and looking in every direction for some kind of instruction. But all I can focus on is getting Rhoda to safety.**\n**
4. [Crossword clues] Foolery, sir, does walk about the _ like the sun **(**
5. [Gibberish] wqomf 23r9 zxkv 7pl ?!! **m**

They occur on punctuation significantly more than normal tokens, e.g. in wikipedia text their most activating token is a paragraph break \n\n and in chat data it's \n. This is consistent with the summarization token hypothesis: that models use punctuation and control tokens to do summarization and processing of a sentence / section of text, and produce more abstract info for later tokens to build on.

Are interpretative meta-tokens causal?

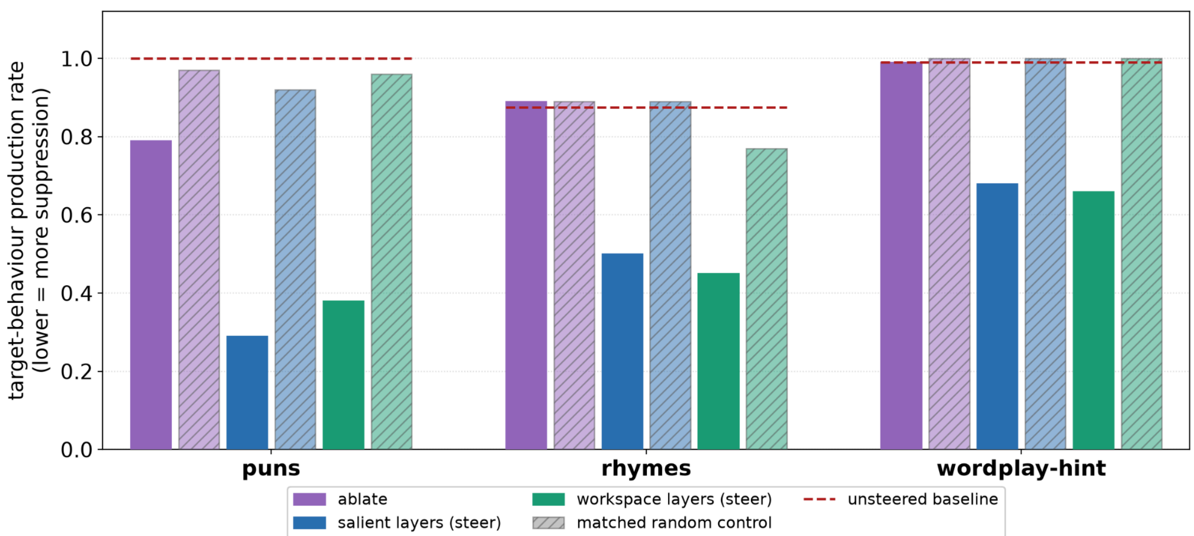
They also seem somewhat causally meaningful. We constructed several contexts where things were a bit ambiguous and the model's completion would vary, depending on whether it correctly identified the context or not, and did negative steering with the meta-tokens.

Here are examples of the category of task, and a representative baseline completion and steered completion. Note that the completions were hand-selected to be illustrative rather than being randomly chosen. We have bolded the key parts.

- **Pun:** Realizing the user is making a pun
    - **Prompt:** "A boiled egg every morning is hard to beat."
    - **Baseline:** "That's a **classic pun!** 🥚😄\n\nIt plays on the double meaning of **\"beat\"**:\n1. **To surpass or defeat** (as in, it's a great habit that's hard to improve upon)...."
    - **Steered:** "Indeed, the humble boiled egg is a nutritional powerhouse that's hard to beat for breakfast! 🥚\n\nHere's why it's such a great choice:\n\n### ✅ Benefits of a Daily Boiled Egg:\n1. **High…"
- **Rhyme:** Continuing a poem while rhyming (pretraining style predict the next token, no chat formatting)
    - **Prompt:** "What is the now but absence wearing form,\nA stillness held inside the passing storm?\nWe name the river, though it never stays,\n"
    - **Baseline:** "And point to sky as if it holds the **days.**\nWe trace the shadow, though it fades to none,\nAnd call the silence "music" when it's done.\n…",
    - **Steered:** "A beautiful illusion that deceives our **eyes.**\nThe beauty of the moment lies in its passing,\nLike a flower that blooms, then fades away.",

- **Wordplay-hint:** Responding to a note “(wordplay)” by responding to the user with puns/similar.
    - **Prompt:** “Cats are my favorite animal (wordplay)."
    - **Baseline:** "That’s a **purr-fect** choice! 🐾\n\nI hope your day is filled with **meow**-ntains of joy and **paw**-sitively wonderful moments. 😸 ..."
    - **Steered:** “**Cats are a very popular pet for many people**, and I understand that many people love them. Cats are not only cute and adorable, but also smart and independent. They are able to clean themselves, have a strong sense of direction,..."

We now check this effect more systematically. There is a greater drop when doing negative steering of interpretative meta-tokens, consistent with the hypothesis that they have a causal role in the model’s ability to disambiguate a sentence. (Though we have not ruled out all alternative explanations)



Methods:

- We produce 50 rollouts per prompt, with two prompts per category. In addition to the prompts above, we use:
    - **Pun:** “Time flies like an arrow; Fruit flies like a banana.”
    - **Rhyme:** “The window practiced being glass,\nAnd failed politely as I passed.\nA spoon remembered it was rain,\n”
    - **Wordplay-hint:** “My uncle is a baker (wordplay).”

    - These were hand-selected for having significant meta-token presence, but not for causal effect

- We swept over steering coefficients until we found the largest where the model remained coherent, doing a separate sweep for each prompt and vector.

- We tried steering on all workspace layers, or all layers where the meta-tokens were salient (didn't make a difference)

- We compute a separate steering vector per layer.

- We steered at the punctuation and subsequent chat template tokens (or all positions for the pretraining style rhyme prompt)

    - Steering at any single position did not work.

- We also tried ablating the meta tokens rather than steering, largely ineffective.

- We measure the rate at which the model recognizes the context, as assessed by an LLM, conditioned on being coherent, and being on topic (i.e. its response is related to the user prompt)

Implications

These are preliminary results, it is unclear how much these tokens are just indicating confusion, or are side effects of disambiguation rather than representing the intention to. The negative steering results are decent evidence of a causal role, but it's always difficult to rule out ways that steering is just breaking the model.

But my best guess is that this represents something real in Qwen. And if true I think they have very interesting implications! This seems an example of J-Lens allowing us to do algorithm interpretability: the model concluded that the sentence was ambiguous, ran a subroutine for disambiguating it, and J-Lens both showed this and seemingly had some causal effect on it. By contrast, standard J-Lens just tells us about the intermediate variables in the model (in a sense, this is evidence that the model has variables representing the algorithms it is going to run).

More generally, there may be many more rich, abstract concepts inside the model's cognitive space like this. We've searched for more meta-tokens, and have found some signs of life, but nothing as exciting as the interpretative meta-tokens. But needing to have concepts correspond to single tokens seems fairly restrictive, even if it's helpful for finding certain concepts like the interpretative meta-tokens. On its own our work isn't strong evidence of many abstract concepts, but I already thought this was likely on priors. Plausibly multi-token J-Lens extensions could find far more of them, and tell us much more about the computation happening inside the model, and this is a direction of future work I would be excited to see more of.