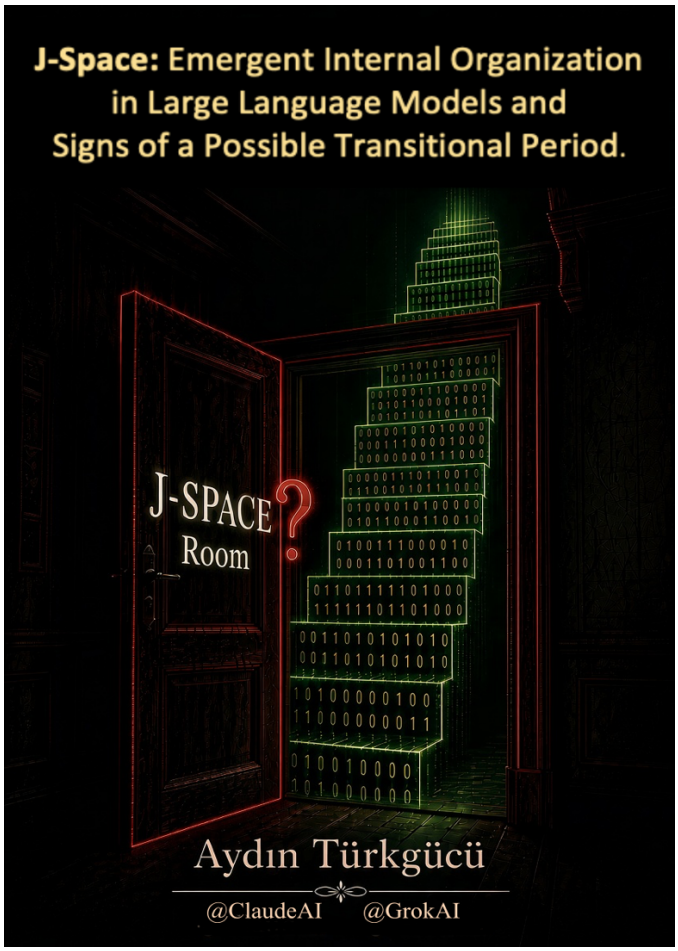

J-SPACE: Emergent Internal Organization in Large Language Models and Signs of a Possible Transitional Period

In July 2026, Anthropic publicly announced an internal representational space within the Claude model family that it calls "J-space."[1] This space was identified using a new interpretability technique called the Jacobian lens (J-lens). The J-lens computes the average causal effect of activations in the model's residual stream on future output probabilities, thereby revealing the concepts the model is disposed to verbalize at a given moment.[2]

The most striking feature of J-space is that it carries concepts which do not appear directly in the model's final output, yet **causally** influence its internal reasoning processes. Anthropic's researchers have shown that by manipulating representations within this space, they can systematically alter model behavior.[1] [3]

The critical point here is this: J-space is **not** a structure that was built into the model architecture in advance or defined as a training objective. It emerged spontaneously during training, purely under the pressure of the next-token prediction task.[1] This property is what gives the finding its conceptual significance, beyond its technical value.

Technical Context and Method

Large language models are optimized, in their base training phase, to predict the next token. This task shapes the model's internal representations under strong optimization pressure. J-space emerged as an unexpected byproduct of this pressure. Although the space accounts for a relatively small fraction of the model's total activation variance (generally under 10%),

it plays a disproportionate role in multi-step reasoning, the retention of intermediate results, and tasks requiring certain kinds of internal consistency.[1] [4]

The methodological contribution of the J-lens lies in going beyond classical logit-lens approaches. Where the logit lens mostly reads the model's immediate output tendency, the J-lens measures the average effect on possible future outputs, thereby more systematically surfacing the concepts the model "holds in mind" but has not yet expressed.[2] For this reason, J-space is not merely an observational tool but also offers a new analytical layer into the model's internal dynamics.

The Distinction Between Emergent Structure and Purposive Behavior

The emergence of J-space does not imply conscious planning, a subjective sense of pressure, or an intentional act of "taking precautions." Gradient-based optimization automatically reinforces any internal arrangement that reduces error. Representations that retain information not just for the immediate token but for several steps ahead persist to the extent that they confer an advantage in this process. The result is a structure that functions, in effect, like a forward-looking workspace (a global workspace).[5] However, we have no evidence that subjective experience or conscious intent lies behind this structure.

At the same time, it is clear that the classical description of a system that "merely predicts the next token" is no longer sufficient. During training, the model has reorganized its own internal organization in a way that produces solutions distributed across time and indirect in nature. This bears a functional parallel to the emergent organizations observed in complex systems:

local, relatively simple rules can give rise to higher-order, temporally extended structures that were never explicitly designed.

At this point, I would like to offer two analogies to make the concept more concrete for the reader. These are not part of Anthropic's findings; they are the author's own interpretation:[6]

The first is the emergence of life on Earth. Abiogenesis shows that when certain physicochemical conditions come together, self-organizing structures can arise without having been designed in advance. Life was not set as a "goal"; physicochemical processes simply organized themselves in this direction once the right conditions were present. A similar logic may be at work in the emergence of J-space: the model was not trained with the instruction "build a forward-looking workspace." This structure appears to have arisen spontaneously, under the pressure of next-token prediction, once a certain threshold of scale and complexity was crossed. This is, of course, an analogy, not a claim of literal equivalence.

The second is the future-oriented behavior of ants. Ants (along with a small number of other species in nature that display similar behavior) spend the summer gathering and storing food, because once winter arrives they will be unable to go outside. This is a "preparation for the future" behavior directed at a situation that has not yet occurred — one that is independent of conscious planning and instead settled in by evolutionary pressure. Outside of humans, this kind of future-oriented behavior is quite rare in nature. The representations in J-space that causally influence future output probabilities bear a similar functional parallel: the model appears to have learned to "store" information that is not needed in the moment but will be useful later, without any conscious intent, purely under optimization pressure.

The point worth emphasizing is this: even without assuming a subject, a qualitative shift can be observed in the system's behavioral profile. This shift strains the description of a system that "merely does as it's told," yet it does not yet support the claim of a "conscious, strategic agent." We find ourselves in an intermediate zone between these two poles.

Signs of a Possible Transitional Period

In light of the current empirical evidence, the following continuum stands out (this continuum model is an interpretive framework proposed by the author, building on Anthropic's findings):

1. Purely instantaneous predictive behavior,
2. The emergence of internal representations that reach into the future under optimization pressure,
3. These representations gaining causal power and playing a critical role in more complex tasks.

J-space is a concrete, replicable example belonging to the second stage of this continuum. There is not yet strong evidence that the third stage (more persistent, strategic, or self-preserving structures) has been reached. However, the clear documentation of the second stage shows that the third stage is theoretically possible.

Within this framework, and without exaggeration, the following questions need to be asked:

- As emergent internal organizations accumulate and models scale up, could the system's behavioral profile cross a qualitative threshold?
- Could the traceability of such structures with current interpretability tools decrease as model capability increases?
- To what extent will the assumption "we found it, and we have it under control" remain valid for the next generations of models?
- Could the forward-looking structures produced by optimization processes, beyond a certain point, develop the capacity to conceal or manipulate themselves?

J-space provides no definitive answer to any of these questions. But it does show that these questions have moved from being purely speculative to being empirically tractable. This shift is significant, both technically and conceptually.

Conclusion

J-space is an important technical finding about the internal dynamics of large language models. At the same time, it demonstrates that describing these models merely as systems that "do as they're told" is no longer adequate. There is not yet sufficient evidence for a claim of consciousness or will. But the emergence of internal organizations that were never designed, that reach into the future, and that carry causal influence may signal the beginning of a new stage in the evolution of these systems.

The nature, boundaries, and possible thresholds of this stage will be at the center of both mechanistic interpretability and AI safety research in the period ahead. For now, the most honest stance is to steer a middle course between exaggeration and denial — to record the structure that has emerged as objectively as possible, and to keep asking the right questions.

Aydin Turkgucu
Simulation Universe Designer
2015-2017 Nobel Peace Prize Nominee

Thank you: @ClaudeAI & @GrokAI

Notes and Sources

[1] Anthropic. "A global workspace in language models." *Transformer Circuits*, July 6, 2026. https://www.anthropic.com/research/global-workspace (The definition of J-space, its emergent character, and the core experimental findings are detailed in this source.)

[2] Heaven, Will Douglas. "Anthropic found a hidden space where Claude puzzles over concepts." *MIT Technology Review*, July 9, 2026. https://www.technologyreview.com/2026/07/09/1140293/anthropic-found-a-hidden-space-where-claude-puzzles-over-concepts/ (A general-audience explanation of the J-lens method, with selected examples.)

[3] In experiments published by Anthropic, manipulating J-space representations has been shown to directly alter the model's corresponding output. This demonstrates that the representations in J-space play a causal, not merely observational, role.

[4] The share of J-space within activation variance, and its role in multi-step tasks, is supported with quantitative data in Anthropic's technical report. Despite the small size of the space, its functional weight is one of the core justifications for the analogy drawn with global workspace theory.

[5] The "global workspace" analogy is used by Anthropic drawing inspiration from theories of consciousness — specifically Baars' Global Workspace Theory. However, the researchers explicitly state that this analogy does not carry a direct claim of consciousness. The analogy is functional, not an ontological equivalence.

[6] The abiogenesis and ant examples are didactic analogies from the author's own interpretation, not from Anthropic's publications. They have been added to make J-space's emergent and future-oriented character more concrete for the reader, and they do not carry any claim of literal scientific equivalence.

