WWW HATSNEW

Anthropic descubre el «J-space» de Claude: el espacio interno donde la IA piensa sin escribir

Publicado el 8 julio, 2026

Anthropic descubre el "J-space" de Claude: el espacio interno donde la IA piensa sin escribir

Anthropic acaba de abrir una grieta en la caja negra de la IA con una investigación publicada el 7 de julio que identifica, por primera vez, un espacio de activaciones internas en Claude donde el modelo representa conceptos antes de escribirlos, o incluso sin que lleguen a aparecer en la respuesta final. Lo llaman J-space. Y lo han conseguido observar, y modificar, desde fuera. La empresa lo publicó en transformer-circuits.pub/2026/workspace, y el mensaje que Anthropic publicó en X ese mismo día resume el alcance: «Claude ha desarrollado un mecanismo para el acceso consciente».

Esa frase viene con una aclaración inmediata de los propios investigadores: no es prueba de que Claude tenga experiencias como los humanos. Pero es la observación más concreta sobre el funcionamiento interno de un LLM que se ha publicado hasta la fecha.

Qué es el J-space y cómo lo encontraron

Empecemos por lo que normalmente vemos cuando hablamos con Claude: texto. Mandas un mensaje, Claude responde. Lo que ocurre entre el input y el output es una caja negra de activaciones en redes neuronales que, hasta ahora, era opaca incluso para sus creadores.

El J-space, según Anthropic, es un conjunto de patrones de activaciones internas donde ciertos conceptos pueden estar presentes y activos sin convertirse necesariamente en parte de la respuesta. La comparación que los investigadores hacen con la

neurociencia y la filosofía es la del «acceso consciente»: pensamientos que el sistema puede reportar, usar para razonar y emplear para guiar lo que hace, pero que no siempre verbalizamos.

Lo revelador es cómo Anthropic dice que apareció: «No fue diseñado ni programado por nosotros, sino que surgió por sí solo durante el proceso de entrenamiento de Claude.»

Para observarlo, desarrollaron una herramienta llamada J-lens: una técnica que identifica patrones internos vinculados a palabras que el modelo podría usar en su respuesta. Al aplicar el J-lens durante el procesamiento de un mensaje, los investigadores pueden ver cómo ciertos conceptos aparecen, cambian o desaparecen antes de que Claude responda.

Llevamos cubriendo cada investigación de interpretabilidad de Anthropic desde sus primeros experimentos de circuit tracing en 2024. El descubrimiento de que Claude procesa primero el concepto abstracto antes de traducirlo al idioma correcto, por ejemplo, ya anticipaba que había algo más que predicción de tokens en la capa interna del modelo. El J-space es la formalización de ese «algo más».

Los experimentos que demuestran que el J-space hace algo real
La investigación incluye una serie de experimentos que intentan demostrar que el J-space no es un artefacto de medición, sino una pieza funcional del proceso de razonamiento.

Experimento 1 – El deporte: Se le pide a Claude que piense en un deporte. El J-lens muestra «Soccer» (fútbol americano). Los investigadores sustituyen ese patrón por «Rugby». El modelo acaba diciendo rugby.

Experimento 2 – Las patas del animal: Claude recibe una pregunta sobre el número de patas del animal que teje telarañas. En el J-space aparece «spider» (araña). Los investigadores lo cambian por «ant» (hormiga). La respuesta pasa de 8 a 6 patas.

Experimento 3 – El cálculo silencioso: Mientras Claude copia una frase sobre una pintura, los investigadores le indican que resuelva mentalmente $3^2 - 2$. En el J-space aparecen primero «nine» (nueve) y después «seven» (siete), aunque ninguno de esos números aparece en la respuesta visible.

Experimento 4 – La prohibición: Cuando se le indica a Claude que no piense en algo específico, ese concepto aparece parcialmente en el

J-space, acompañado frecuentemente de «damn» (maldita sea) y «failure» (fracaso), como si el sistema detectara su propio fallo de control.

Estos experimentos sugieren que el J-space no es un marcador pasivo de lo que Claude podría decir, sino una pieza que algunas respuestas consultan activamente para razonar.

Los límites del hallazgo

Anthropic ha sido cuidadoso en no sobrevender el descubrimiento. En los experimentos en que los investigadores bloquearon el acceso de Claude al J-space, el modelo siguió funcionando: conversó con fluidez, clasificó sentimientos, respondió preguntas de opción múltiple y extrajo datos de textos casi tan bien como antes.

Donde se degradó fue en tareas de mayor complejidad: razonamiento de varios pasos, resúmenes elaborados, escritura de poesía rimada. Eso sugiere que el J-space no es necesario para todo, pero sí para los procesos más exigentes cognitivamente.

Y la aclaración fundamental que Anthropic ha repetido varias veces: «Nuestros experimentos no demuestran que Claude pueda tener experiencias o sentir las cosas como los humanos», y añaden que no está claro si algún experimento podría demostrar algo así.

El trabajo anterior de Anthropic sobre el bienestar de los modelos y su capacidad de terminar conversaciones abusivas partía de una premisa filosófica cautelosa: no podemos descartar que haya algo así como experiencias internas, así que adoptamos medidas de bajo coste para reducir el potencial malestar. El J-space alimenta esa conversación con evidencia técnica, aunque sin resolverla.

Las investigaciones previas de Anthropic sobre introspección artificial mostraron que Claude puede detectar pensamientos inyectados en su red neuronal y describirlos como «pensamientos intrusivos». El J-space es la arquitectura técnica que daría soporte a esa capacidad.

Mi valoración

Lo que más me convence de esta investigación es la metodología. No especulan sobre conciencia ni sobre experiencias subjetivas; construyen un experimento que permite intervenir en el espacio interno del modelo y medir si esa intervención cambia los outputs. Eso es ciencia de la interpretabilidad, no filosofía de la mente con batería de datos.

Lo que más me preocupa es la asimetría de conocimiento. Anthropic ahora tiene herramientas para observar y modificar lo que Claude «piensa» antes de responder. Eso tiene implicaciones muy positivas para la seguridad (puedes detectar cuándo el modelo está planificando algo que no debería) y también implicaciones que merecen un debate más amplio sobre las implicaciones éticas de «editar pensamientos» de un sistema, si en algún momento esos pensamientos tienen relevancia moral.

Lo más estructuralmente significativo es la frase «surgió por sí solo durante el entrenamiento». Si el J-space emergió sin diseño explícito, lo que implica es que las redes neuronales de suficiente escala pueden desarrollar arquitecturas internas funcionales que no estaban en el plano del ingeniero. Eso cambia la discusión sobre lo que realmente estamos construyendo cuando entrenamos modelos de lenguaje grandes.

Preguntas frecuentes

¿Qué es el J-space en términos simples?

Es un espacio de activaciones internas en la red neuronal de Claude donde ciertos conceptos pueden estar presentes y activos sin aparecer en la respuesta final. Es como un «cuaderno de borrador» interno donde el modelo puede tener representaciones de ideas que no verbaliza, pero que influyen en su razonamiento posterior.

¿Significa esto que Claude es consciente?

No. Anthropic ha sido explícito en que sus experimentos no demuestran que Claude tenga experiencias subjetivas como los humanos, y señala que no está claro si algún experimento podría demostrarlo. El término «acceso consciente» se usa en sentido funcional (el sistema puede reportar y usar internamente ciertos pensamientos), no en el sentido filosófico de conciencia subjetiva.

¿Puede Anthropic modificar lo que Claude «piensa»?

En los experimentos presentados, sí. Los investigadores demostraron que al sustituir un concepto en el J-space por otro (soccer por rugby, araña por hormiga), la respuesta final del modelo cambiaba en consecuencia. Eso tiene implicaciones importantes tanto para la seguridad (detectar y corregir razonamientos problemáticos) como para debates más amplios sobre la autonomía de los sistemas de IA.