KAMAY Revista de Filosofía
Vol. I, N° 1, 2026 | pp. 151-183

¿Entienden los LLMs lo que producen? Producto, proceso y agencia en la atribución de entendimiento superficial

Do LLMs Understand what they Produce? Product, Process, and Agency in the Attribution of Superficial Understanding

Jorge Sebastián Polo Núñez
Universidad Nacional de San Antonio Abad del Cusco

Resumen: Este artículo examina si los modelos de lenguaje grandes (LLMs) "entienden" lo que producen. Para evitar inferencias apresuradas desde la calidad del texto hacia estados mentales, se adopta la distinción de Boisseau entre imitación como conducta y como estatus del producto, caracterizando a los LLMs como fábricas de imitaciones lingüísticas. Sobre ese marco, se articulan dos apoyos independientes: (i) la tesis de Borg de que los outputs pueden portar contenido semántico a nivel de tipo lingüístico mediante deferencia semántica, sin requerir intencionalidad original; y (ii) la propuesta de Piantadosi y Hill según la cual los estados internos de los LLMs pueden instanciar aspectos del significado entendido como rol conceptual. Se añade una discusión sobre el grounding referencial limitado (Coelho Mollo y Millière). Se argumenta que la conjunción de estos tres apoyos —significado derivado del producto, estructura conceptual parcial del proceso, y anclaje referencial indirecto— autoriza a proponer una categoría intermedia, aquí denominada entendimiento superficial, definida por las condiciones C1–C4. Frente al argumento de Hattangadi y Schoubye sobre la carencia de significado literal, se ofrece una doble respuesta: semántica, al mostrar que su argumento anti-CRS no alcanza a la versión moderada aquí adoptada; y epistemológica, al mostrar con Collins y Evans que la evaluación del éxito imitativo depende de la competencia del juez. Finalmente, se discute la relación de la tesis con el bibliotechnism de Lederman y Mahowald, y se reconoce el problema de la referencia novedosa como un límite abierto.
Palabras clave: inteligencia artificial; procesamiento del lenguaje natural; semántica; epistemología; evaluación; comunicación

Abstract: This article examines whether large language models (LLMs) "understand" what they produce. To avoid hasty inferences from text quality to mental states, we adopt Boisseau's distinction between imitation as behavior and as a product status, characterizing LLMs as factories of linguistic imitations. Within this framework, two independent arguments are articulated: (i) Borg's thesis that outputs can carry semantic content at the linguistic type level through semantic

content at the linguistic type level through semantic deferral, without requiring original intentionality; and (ii) Piantadosi and Hill's proposal that the internal states of LLMs can instantiate aspects of meaning understood as conceptual roles. A discussion of limited referential grounding (Coelho Mollo and Millière) is added. It is argued that the combination of these three supports—product-derived meaning, partial conceptual structure of the process, and indirect referential grounding—justifies proposing an intermediate category, here termed superficial understanding, defined by conditions C1–C4. In response to Hattangadi and Schoubye's argument regarding the lack of literal meaning, a twofold response is offered: semantic, by showing that their anti-CRS argument does not hold against the moderate version adopted here; and epistemological, by demonstrating, along with Collins and Evans, that the evaluation of imitative success depends on the judge's competence. Finally, the thesis's relationship to Lederman and Mahowald's bibliotechnism is discussed, and the problem of novel reference is acknowledged as an open question.
Keywords: artificial intelligence; natural language processing; semantics; epistemology; evaluation; communication

Introducción

Imagina que llegan a tu bandeja dos cartas impresas, las dos son sobrias, coherentes, sensibles al contexto, con un tono creíble y una estructura impecable; te dicen que una la escribió una persona y la otra un modelo de lenguaje. Si no puedes distinguirlas solo por el texto, ¿qué estás evaluando exactamente cuándo preguntas si el modelo entiende? La posición inmediata frente a ello es tomar el producto final, la carta, como una ventana directa a la mente que supuestamente la produjo. Pero esa inferencia es justamente lo que el debate contemporáneo vuelve una y otra vez problemático: el término entendimiento tiende a operar como un bloque que mezcla al menos tres cosas distintas: (i) el estatus semántico del texto, (ii) las capacidades que explican que ese texto sea posible, y (iii) la agencia epistémica que permitiría responsabilizar a un sujeto por lo dicho.

Para desarmar ese bloque conviene empezar por un retrato mínimo de qué hace un LLM. La mayoría de los modelos actuales se basan en la arquitectura transformer, cuyo mecanismo distintivo es la autoatención: para generar o interpretar una secuencia, el sistema pondera qué partes del contexto son relevantes para cada token y construye representaciones dependientes del contexto, no meras

asociaciones fijas palabra a palabra (Ferraris et al., 2025). Muchos sistemas conversacionales de uso común adoptan un diseño decoder-only y se entrenan con una meta clara: predecir el siguiente token de una secuencia en función del contexto previo, esta receta técnica, que parece modesta, permite un rendimiento sorprendente en tareas como generación de texto, traducción o preguntas y respuestas; al mismo tiempo, trae limitaciones que son filosóficamente reveladoras, como la tendencia a producir enunciados plausibles pero falsos, y la necesidad de apoyos externos como la recuperación de información para reducir ese problema (Ferraris et al., 2025).

Dicho en términos más simples: tenemos un sistema capaz de producir lenguaje muy convincente sin que, por diseño, sea obvio que su operación esté orientada a la verdad, a la referencia o a compromisos epistémicos del modo en que lo está un hablante humano. En este punto reaparecen los experimentos mentales de Turing y Searle; Borg (2025) recuerda que buena parte de la discusión sobre inteligencia artificial se ha estructurado alrededor de dos experimentos mentales: el test de imitación de Turing, que propone un criterio conductual de indistinguibilidad conversacional, y el cuarto chino de Searle, que busca mostrar que el éxito lingüístico puede lograrse mediante manipulación formal sin comprensión. Si el test de Turing nos lleva a evaluar el fenómeno por el producto observable, el cuarto chino subraya que el producto puede ser compatible con procesos que no sería adecuado señalar con la etiqueta de entendimiento. Al evaluar los LLMs, las diferencias se intensifican porque su desempeño hace natural la lectura en el marco del test de Turing, mientras que sus fallos y opacidades reactivan la sospecha searleana asociada al cuarto chino.

El aporte de Boisseau (2024) es que esta disputa se mantiene irresoluble, en parte, por una ambigüedad conceptual previa relacionad con la noción de imitación. Se habla de imitación como si fuera una sola cosa, cuando en realidad se usan dos nociones distintas: una es la imitación como conducta, es decir, un agente se

comporta como otro, con un desempeño dirigido y un criterio de éxito ligado a "pasar por" alguien. La otra es la imitación como estatus del producto, cuando algo es una imitación de otra cosa porque fue producido a partir de un modelo, como una falsificación o un sustituto. La tesis central de Boisseau es que los LLMs encajan mejor en esta segunda familia: no son, en sentido estricto, agentes que imitan, sino sistemas que fabrican productos con estatus de imitación del habla humana. Esa categoría, que Boisseau denomina imitation manufacturing, permite describir con más precisión por qué el output puede ser tan persuasivo sin que por eso sea obligatorio atribuir comprensión al sistema. Boisseau misma adopta una posición de no-entendimiento, pero su marco conceptual deja abierto un espacio que este artículo aprovecha para construir una tesis intermedia.

Este artículo defiende una tesis intermedia, construida a partir de esa distinción, con Boisseau (2024), sostengo que es metodológicamente más claro pensar muchos LLMs como fábricas de imitaciones lingüísticas, es decir, como mecanismos de producción de textos que pueden circular como si fueran habla humana. Sin embargo, esta caracterización no pretende reducirlos a un mero loro estocástico; para matizar la imagen del entendimiento y dar lugar al entendimiento superficial incorporo dos apoyos. Primero, Borg (2025) ofrece un argumento para tomar con precisión el significado del producto: aunque el sistema no posea intencionalidad original ni agencia consciente, los textos que produce pueden ser genuinamente significativos en un sentido derivado, por su inserción en prácticas lingüísticas humanas y por el hecho de que operan con signos del lenguaje natural ya cargados de normas de uso. Segundo, Piantadosi y Hill (2022) proponen que una parte del significado puede entenderse como rol conceptual, determinado por relaciones entre estados internos. Desde esa perspectiva, es plausible que haya estructura semántica parcial en los LLMs, aun cuando falten componentes que harían su semántica plenamente humana, como metas propias, percepción situada o control normativo robusto. La conclusión que quiero defender es qué: lo que estas tres líneas de

trabajo autorizan —cada una desde un ángulo distinto y sin que ninguna proponga esta categoría por sí sola— es la construcción de una noción de entendimiento superficial. Defino entendimiento superficial como un nivel de éxito lingüístico que satisface tres condiciones positivas: suficiencia semántico-pragmática del producto, estructura interna parcialmente semántica, y anclaje referencial limitado en ciertos casos; pero que no satisface las condiciones de comprensión robusta: agencia epistémica, control normativo, orientación estable a la verdad.

1. La Distinción de Boisseau sobre la Imitación

Una parte importante de la confusión en la discusión sobre LLMs se debe a que el término imitación se utiliza como si nombrara un fenómeno único, pero en el uso ordinario y también en la literatura filosófica, el mismo término sirve para hablar de cosas distintas. Boisseau (2024) sostiene que esta ambigüedad no es un detalle terminológico, sino un punto metodológico importante, si no desambiguamos, el debate sobre entendimiento corre el riesgo de depender de transiciones entre sentidos diferentes, y con ello de inferencias injustificadas. La propuesta de esta sección es reconstruir la distinción que plantea la autora y mostrar por qué, aplicada a LLMs, permite separar con mayor precisión lo que pertenece al producto lingüístico de lo que pertenece al proceso que lo genera.

1.1. Dos sentidos de imitación

En un primer sentido, imitar es una forma de conducta: un agente imita cuando desempeña una acción a la manera de otro, el caso paradigmático es la imitación de una persona concreta al adoptar su estilo, sus gestos, su acento, su modo de argumentar. Este tipo de imitación tiene una estructura práctica reconocible porque está orientada por un objetivo, presupone selección de rasgos relevantes y suele involucrar ajuste a partir de retroalimentación, de esa manera el imitador intenta pasar por el otro, o al menos

producir en el público la impresión de estar ante el otro. Incluso cuando la motivación no es engañar, hay un componente de control, ya que la conducta se regula por un criterio de éxito que remite al modelo imitado (Boisseau, 2024).

Lo crucial para el argumento posterior es que este sentido de imitación es afín a la agencia, porque la imitación conductual no es solo semejanza entre dos productos, sino más bien, es una actividad que se organiza alrededor de una meta y se corrige en función de esa meta. Esto no implica que todo imitador deba poseer reflexión explícita o deliberación sofisticada, pero sí sugiere una clase de desempeño dirigido que se describe mal si se reduce a la mera producción de una salida parecida. Un ejemplo que ayuda a fijar el contraste es el siguiente: un estudiante puede imitar el estilo de un profesor en un ensayo, copiando su estructura de exposición, su vocabulario y su forma de atender las participaciones de otros estudiantes. Si el resultado se parece al profesor, decimos que lo ha imitado, pero el rasgo central no es solo el texto final, sino que el estudiante intentó producir ese efecto y ajustó el desempeño para lograrlo, en este sentido, la imitación es primariamente un tipo de hacer por un agente.

En un segundo sentido, hablar de imitación es hablar del estatus de un producto, es decir algo es una imitación de otra cosa cuando tiene una relación genealógica con un modelo y se parece a él de manera no accidental, en estos casos el paradigma son las falsificaciones o sustitutos, como el cuero sintético, billetes falsos, réplicas de una obra de arte, incluso una clase de productos diseñado para parecerse a otros. En este sentido no es necesario postular una conducta en curso ni un agente que esté actuando como alguien, el foco está en el artefacto producido y en la manera en que su semejanza está explicada por el modo de producción (Boisseau, 2024). Este segundo sentido introduce, además, una dimensión normativamente relevante: la confundibilidad, porque muchas imitaciones en este sentido se caracterizan por poder ser tomadas por el original en ciertos contextos y justamente por ello

puede ser buscada para engañar, pero también puede ser funcional sin engaño. Un material puede ser imitación de madera porque cumple ciertas expectativas perceptivas y prácticas, aunque se venda explícitamente como sustituto, lo decisivo no es el engaño, sino el tipo de semejanza y su relación con el modelo.

Si dos estudiantes escriben la misma frase aprendida de un manual, hay reproducción del enunciado, pero no necesariamente imitación entre ellos, la imitación como estatus requiere que el parecido esté vinculado a un procedimiento de copia, modelado o producción en referencia a un original (Boisseau, 2024). Esta cláusula de no accidentalidad cumple un papel importante cuando pasamos a LLMs, porque permite preguntar de dónde proviene el parecido entre outputs y habla humana.

1.2. Imitation Manufacturing como Categoría para LLMs

La tesis más característica de Boisseau es que describir a los LLMs como imitadores en sentido conductual es ambigua, porque puede sonar natural decir que un modelo imita a los humanos porque produce texto con tono humano, sin embargo, el sentido conductual introduce elementos que no están claramente presentes en el caso. Un LLM no organiza su desempeño por la meta de pasar por un humano del mismo modo en que lo hace un imitador humano, el criterio de éxito de la generación no es, desde el punto de vista del sistema, parecerse a una persona como finalidad adoptada, de esta manera su operación se caracteriza mejor como un procedimiento que, dadas ciertas condiciones de entrada, produce continuaciones que son estadísticamente adecuadas al contexto de entrenamiento. En esa medida, lo que se obtiene es un producto con propiedades de semejanza, pero la semejanza se explica por el entrenamiento sobre datos humanos y por el diseño del sistema, no por una intención propia de imitar a un sujeto (Boisseau, 2024).

Para evitar esta confusión, Boisseau propone hablar de imitation manufacturing, la idea no es negar que haya imitación, sino ubicarla donde corresponde, ya que el modelo funciona como un

sistema de producción que fabrica salidas lingüísticas con estatus de imitaciones del habla humana. Es una tesis de localización conceptual: lo que imita es el output, en el sentido de estatus del producto, y el sistema es el mecanismo que produce ese tipo de outputs. Una analogía útil es la del taller que fabrica réplicas, en el que se produce objetos que pueden pasar por originales bajo ciertas condiciones, y ese hecho es inteligible sin postular que el taller actúe como el artesano original.

Este cambio de categoría permite una formulación más estable de la pregunta por el entendimiento, porque en lugar de preguntar sin más si el modelo entiende porque su output parece humano, podemos preguntar qué propiedades del proceso estarían justificadas por el hecho de que produce outputs con estatus de imitación de habla humana. La respuesta no es inmediata ni única, pero el marco impide una inferencia precipitada, ya que habilita que si el objetivo de la imitación es el producto eso no implica que el proceso también sea imitado en sentido estricto, del parecido del producto no se sigue automáticamente la posesión, por el sistema, de las capacidades agenciales asociadas a la imitación conductual.

La utilidad de esta distinción se ve en un error recurrente del debate: inferir propiedades psicológicas o epistémicas del sistema a partir de propiedades estilísticas o semánticas del texto generado. Cuando un LLM produce una carta con tono adecuado, buena estructura argumentativa, es tentador concluir que el sistema ha comprendido lo que escribe, por ello la distinción de Boisseau fuerza a distinguir dos evaluaciones distintas: una evaluación del producto pregunta si el texto es interpretable, coherente, apropiado, y si puede circular como pieza de habla humana; y otra evaluación del proceso pregunta por el tipo de capacidades que explican esa producción y por si esas capacidades incluyen compromiso con normas epistémicas, control de verdad, o agencia semántica robusta.

Aquí aparece el punto que guiará el resto del artículo porque la discusión suele tratar producto y producción como un bloque unificado, se toma el hecho de que el producto exhibe rasgos que asociamos al entendimiento humano y se proyecta esos rasgos al sistema, por ello la distinción de Boisseau no resuelve por sí sola la cuestión de si hay entendimiento en algún sentido, pero reduce la confusión conceptual. Permite decir, con mayor precisión, que lo primero que observamos es un artefacto textual con estatus de imitación, y que la carga de la prueba está en mostrar qué se sigue de ello acerca del proceso generativo, sin asumir que deberíamos también de analogar el proceso dado el producto imitado. Con ello se da un espacio para considerar los productos de los LLM sin asumir el entendimiento, pero aun faltarían apoyos para poblar ese espacio de forma distinta al entendimiento humano para ello, a continuación, veremos como Borg (2025) ofrece razones para afirmar que el output puede ser semánticamente significativo sin exigir intencionalidad original, por su parte Piantadosi y Hill ofrecen razones para pensar que el sistema puede albergar estructura semántica parcial si el significado se entiende como rol conceptual. La distinción de Boisseau (2024) es el marco que permite articular ambos apoyos sin colapsar la tesis principal: que el caso de los LLMs se entiende mejor, por ahora, como producción de imitaciones lingüísticas y, en consecuencia, como un candidato a entendimiento superficial antes que a comprensión robusta.

2. Turing y el Cuarto Chino como Disputa sobre Criterios: Producto y Proceso

El marco clásico en el que suele caer la discusión sobre LLMs puede entenderse como una disputa sobre criterios, es decir sobre qué cuenta como evidencia de entendimiento, qué tipo de evidencia es relevante y qué tipo de inferencia se permite desde un desempeño lingüístico hacia una conclusión sobre comprensión. En este sentido, la oposición entre el test de Turing y el cuarto chino de Searle no es solo un choque de intuiciones; es un desacuerdo sobre si el fenómeno debe evaluarse primariamente por el producto observable

o por el proceso subyacente, frente a ello Borg (2025) reconstruye este punto con claridad al mostrar que el debate contemporáneo hereda la tensión original sin haber estabilizado qué se está midiendo. La estrategia de Turing consistía en desplazar la pregunta metafísica sobre si una máquina puede pensar hacia una prueba práctica, el criterio es conductual: si, en un intercambio conversacional suficientemente rico, un interrogador no logra distinguir a la máquina de un humano, la máquina pasa (Borg, 2025). En el contexto actual, esto se traduce fácilmente al caso de cartas o ensayos: si el texto final es indistinguible del humano para lectores competentes, parecería legítimo hablar de comprensión, o al menos de una capacidad lingüística del mismo tipo.

Pero en el propio uso que Turing hace de la idea de imitación hay una ambivalencia que anticipa la discusión contemporánea si tomamos el ejemplo del examen oral como método para discriminar entre quien entiende una materia y quien la repite de memoria, de manera mecánica. La idea de responder como loro, sugiere que la imitación puede ser precisamente lo contrario del entendimiento (Boisseau, 2024).

El mismo vocabulario, entonces, sirve para argumentos distintos: por un lado, la indistinguibilidad conversacional como signo positivo de inteligencia; por otro, la repetición competente como contraste con la comprensión genuina. Lo que falta, y lo que Boisseau enfatiza, es un marco conceptual que permita decidir qué se sigue de una semejanza del producto sin colapsarla en una tesis psicológica sobre el proceso (Boisseau, 2024).

En el caso del cuarto chino se radicaliza la sospecha contra las inferencias basadas solo en desempeño, la intuición central es que puede haber un sistema que produzca respuestas perfectamente adecuadas en un idioma sin entender ese idioma, siempre que manipule símbolos de acuerdo con reglas puramente formales (Borg, 2025). El objetivo no es negar que haya regularidad en la conducta, sino bloquear el salto desde regularidad conductual a atribución de

estados semánticos y comprensión. El cuarto chino expresa de manera extrema un principio metodológico que reaparece en evaluaciones contemporáneas de LLMs: el producto final puede ser compatible con mecanismos que no merecen, sin más, el vocabulario de la agencia epistémica.

Si un sistema puede producir una carta impecable sin ser responsable de lo que afirma, entonces el texto no funciona como testimonio en sentido fuerte, aunque pueda funcionar como instrumento retórico o como reorganizador de información. Esto no obliga a adoptar el escepticismo total, pero sí obliga a separar cuidadosamente la evaluación del producto de la evaluación del proceso. En este punto, la distinción de Boisseau adquiere su motivación, el debate Turing–Searle tiende a tratar el fenómeno como un bloque o se confía en el output, o se sospecha de él, mientras que Boisseau propone el vocabulario que permite tratarlo como un sistema de producción de imitaciones del habla humana, donde el estatus del producto no decide por sí mismo la atribución de entendimiento al sistema. Con esta preparación, podemos pasar a dos apoyos que complejizan el entendimiento superficial con Borg sobre significado derivado y Piantadosi y Hill sobre rol conceptual.

3. Significado sin Intencionalidad Original

El aporte de Borg (2025) puede leerse como un intento de salir de la dicotomía, en lugar de elegir entre si pasa el test, entonces entiende y podría pasar sin entender, se pregunta qué condiciones suelen exigir los escépticos para admitir significado y si esas condiciones realmente son necesarias, el movimiento obliga a distinguir tipos de atribución: significado del texto, comprensión del sistema y agencia del productor.

Borg reconoce que la tensión Turing y Searle sigue vigente y que, en efecto, un criterio conductual por sí solo no clausura la cuestión del significado. El cuarto chino funciona como advertencia contra una inferencia demasiado rápida desde outputs hacia estados

internos, también insiste en que la intencionalidad original, entendida como la capacidad de un sistema para tener estados intencionales propios, no es algo que debamos atribuir ligeramente a LLMs en su forma actual (Borg, 2025).

Esto conecta directamente con la preocupación de Boisseau: describir al modelo como un agente que imita puede arrastrar supuestos sobre agencia e intención que no han sido establecidos. La contribución decisiva de Borg es separar el problema del significado del problema de la intencionalidad original, parte de la idea de intencionalidad derivada, asociada a Searle, para sostener que un sistema puede manipular símbolos que son significativos porque su significado está fijado por prácticas humanas, aunque el sistema mismo no sea un sujeto con intención comunicativa propia.

Más específicamente, Borg propone un modelo de deferencia semántica a nivel de tipo: los signos que los LLMs manipulan son tokens de tipos lingüísticos cuyas condiciones de significado ya están fijadas por convenciones de las comunidades humanas. Si se adopta una semántica a nivel de tipo (*A-style semantics*), donde el contenido del tipo no depende de la intención actual del hablante sino de la convención, entonces los outputs del LLM expresan contenido semántico a nivel de tipo por el solo hecho de instanciar correctamente esos tipos (Borg, 2025).

Esto es independiente de que el sistema posea intenciones comunicativas o estados mentales propios. Es crucial notar que Borg separa explícitamente tres niveles de atribución: (a) que los outputs sean semánticamente significativos, lo cual defiende; (b) que los estados internos representen propiedades semánticas qua semánticas, lo cual considera una cuestión abierta; y (c) que el sistema entienda o asevere, lo cual niega. En palabras de Borg: “we should deny that LLMs are in the business of asserting the content expressed by the sentences they produce” (2025). El aporte de Borg para el presente argumento se limita, pues, al nivel (a): la significatividad del producto.

La pregunta por el entendimiento del sistema requiere apoyos adicionales. Los LLMs operan precisamente sobre signos del lenguaje natural; sus outputs son materiales que ya pertenecen a un espacio normativo y social de interpretación, en este sentido, no es extraño que las cartas producidas por LLMs sean semánticamente evaluables, incluso cuando el modelo no sea un agente responsable. Este punto encaja la fabricación de imitaciones y el entendimiento superficial, si el output es un producto que circula como habla humana, su estatus de imitación puede incluir, precisamente, su aptitud para ser interpretado como portador de contenido, la carga semántica no prueba agencia sino prueba que el producto pertenece a un régimen de interpretación comunicativa que ya existe. Borg ayuda a defender, entonces, una versión de la siguiente tesis: el texto puede ser significativo sin que el sistema tenga que contar como alguien que quiere decir algo —tesis indispensable para la noción de entendimiento superficial que busco defender más adelante.

Conviene explicitar el marco semántico en juego, ya que Borg opera desde una semántica minimal y composicional, donde el contenido a nivel de tipo se determina por convención lingüística (Borg, 2004). Piantadosi y Hill operan desde la semántica de rol conceptual (CRS), donde el contenido se determina por relaciones entre representaciones (Block, 1986; Harman, 1999). Estos dos marcos no son idénticos, pero son compatibles en el punto que aquí importa: ambos permiten atribuir contenido sin exigir intencionalidad original o referencia directa como condición necesaria. La noción de "estructura semántica parcial" que empleo en este artículo remite específicamente a CRS: designa la presencia de patrones estables de transición inferencial entre estados internos del modelo, no una asignación composicional de valores de verdad a fórmulas.

4. Rol Conceptual y Significado sin Referencia como Requisito

Si Borg fortalece la idea de que el producto puede ser significativo sin intencionalidad original, Piantadosi y Hill (2022) fortalecen otra pieza: la posibilidad de que haya estructura semántica parcial en el proceso, sin que ello equivalga a comprensión robusta, su argumento responde a una crítica frecuente, según la cual los LLMs no pueden tener significado porque carecen de referencia o de anclaje en el mundo, frente a ellas sostienen que el significado no debe identificarse sin más con referencia y que hay buenas razones, desde filosofía del lenguaje y desde ciencia cognitiva, para entenderlo como rol conceptual. La idea del rol conceptual, tal como la desarrollan Block (1986) y Harman (1987, 1999) —y como la retoman Piantadosi y Hill—, es que el contenido de una representación se individúa por su posición funcional en una red de transiciones inferenciales. En este marco, conocido como Conceptual Role Semantics (CRS), lo que fija el significado de un estado mental no es solamente su relación causal con un referente externo, sino el conjunto de relaciones que mantiene con otros estados representacionales: qué inferencias habilita, de qué otras representaciones se sigue, y con cuáles es incompatible. Piantadosi y Hill adoptan una versión moderada de CRS: no niegan que la referencia sea relevante para el significado, sino que la tratan como un aspecto más del rol conceptual total, no como un prerrequisito sin el cual no puede haber contenido alguno (Piantadosi & Hill, 2022).

El argumento de Piantadosi y Hill no es puramente conceptual, señalan evidencia empírica de que las representaciones vectoriales de los LLMs exhiben geometrías isomórficas con la estructura conceptual humana: las relaciones de similitud entre vectores de palabras correlacionan significativamente con medidas representacionales de neuroimagen (fMRI), y los modelos resuelven tareas que presuponen relaciones conceptuales —analogías, paráfrasis, detección de sinonimia y antonimia, resolución de esquemas Winograd— que difícilmente se explicarían si los vectores fueran meros registros de co-ocurrencia sin estructura conceptual (Piantadosi & Hill, 2022).

Esto no prueba comprensión, pero sí vuelve plausible que haya una instanciación parcial de roles conceptuales en el proceso interno del modelo. Piantadosi y Hill (2022) recuerdan que existen expresiones que parecen significativas aunque no tengan un referente claro o aunque su referencia sea problemática, porque conceptos abstractos, ficciones, imposibles o términos vacíos muestran que la referencia no puede ser tratada como requisito universal del significado, el punto no es negar la importancia de la referencia en muchas prácticas lingüísticas, sino resistir una conclusión fuerte que condena la ausencia de referencia directa a un sistema a la carencia total de contenido.

En el marco de este artículo, Piantadosi y Hill presentan un argumento que hace más difícil reducir a los LLMs a una pura imitación vacía, entendida como secuencias sin estructura conceptual, ya que, si el rol conceptual es una vía legítima para hablar de contenido, entonces es plausible que los LLMs alberguen al menos fragmentos de estructura semántica, en el sentido de patrones internos que se comportan como portadores de relaciones conceptuales parciales. Al mismo tiempo, esta plausibilidad no obliga a abandonar el marco de Boisseau, porque que haya estructura semántica parcial no equivale a que el sistema sea un agente que imita en sentido conductual ni a que posea comprensión robusta, Piantadosi y Hill reconocen que un entendimiento plenamente humano estaría ligado a agencia, metas, interacción rica con el entorno y control normativo más estable, en ausencia de esos rasgos, lo que resulta defendible es una atribución limitada, compatible con la idea de fábrica de imitaciones, en el cual el sistema puede sostener regularidades semánticas en su funcionamiento, pero esas regularidades todavía no constituyen el tipo de agencia epistémica que buscamos cuando evaluamos testimonio, responsabilidad o justificación.

Con Borg sobre significado derivado del producto y Piantadosi y Hill sobre rol conceptual parcial del proceso, es razonable hablar de entendimiento superficial en LLMs, siempre que esa expresión se

use para marcar un nivel intermedio, ese nivel reconoce contenido y estructura, pero niega el salto desde el éxito de la imitación del producto a la comprensión robusta del LLMs. Sin embargo, incluso si aceptamos que puede haber estructura conceptual interna y significado derivado del producto, queda abierto un punto clásico: si ese contenido está, de algún modo, anclado en el mundo. La discusión sobre el grounding reaparece aquí como una prueba para cualquier tesis intermedia, porque exige precisar si el entendimiento superficial es solo un efecto de circulación social del lenguaje o si incluye alguna forma limitada de referencia extra lingüística.

5. El Problema del Grounding de los LLMs

Una de las objeciones más recurrentes contra cualquier atribución de entendimiento a los LLMs consiste en apelar al grounding, es decir, si el sistema se entrena exclusivamente con texto, parecería que sus estados internos y sus outputs carecen de anclaje extra lingüístico y, por tanto, no pueden ser sobre el mundo de un modo no parasitario de la interpretación humana. Coelho Mollo y Millière (2025) reformulan esta inquietud como el Vector Grounding Problem: el problema de si los vectores que estructuran las representaciones internas de un LLM pueden sustentar significado intrínseco y referencia independientemente del sentido que los usuarios proyectamos al interactuar con sus salidas. El interés de su propuesta, para el marco que vengo defendiendo, es que desplaza la cuestión desde un escepticismo global, es decir sin sensores, no hay significado, hacia una pregunta más específica, centrada en qué tipo de grounding es relevante y qué condiciones deberían cumplirse para que una atribución semántica limitada sea metodológicamente defendible. La primera contribución consiste en distinguir grounding referencial de otras nociones afines, ya que no todo lo que se llama grounding resuelve el mismo problema, porque puede hablarse de grounding como conexión con percepción y acción, como estabilidad pragmática en uso social, o como simple articulación relacional entre signos. Pero, según Coelho Mollo y

Millière (2025), lo decisivo para responder al desafío escéptico es el grounding referencial, entendido como la conexión entre una representación y su referente en el mundo; esta distinción es crucial porque permite admitir que un sistema pueda carecer de conexión con el mundo fuerte y, aun así, estar en posición de sostener algún grado de referencia, evitando la transición automática desde, los LLMs no tienen sensorimotricidad a no hay contenido alguno en sus respuestas.

El núcleo del argumento es el grounding referencial se logra cuando los estados internos de un sistema satisfacen dos condiciones, la primera, es una condición causal informacional: debe haber una relación apropiada, aunque no necesariamente directa, entre ciertos patrones representacionales y regularidades del mundo; la segunda es una condición histórico funcional: esos estados deben haber sido seleccionados, en este caso ya sea por entrenamiento y evaluación, para portar esa información, de modo que su papel en el sistema pueda caracterizarse normativamente en términos de acierto y error. La tesis defendida es que los LLMs pueden, al menos en principio, satisfacer ambas condiciones, la conexión causal con el mundo puede estar mediada por el hecho de que los datos textuales son productos de agentes humanos situados que describen, registran y corrigen su trato con la realidad; y la dimensión selectiva puede introducirse mediante procedimientos de entrenamiento y ajuste que favorecen sistemáticamente outputs que preservan o recuperan información correcta bajo criterios dependientes del mundo (Coelho Mollo & Millière, 2025).

Este aporte refuerza la defensa del entendimiento superficial, Borg señalaba que el producto puede ser semánticamente significativo de manera derivada por su inserción en prácticas humanas; con Piantadosi y Hill, que el proceso puede albergar estructura conceptual parcial si el significado se entiende como rol conceptual. El argumento sobre grounding añade que no solo el producto hereda normas de uso, y no solo el proceso exhibe regularidades inferenciales internas, sino que también es plausible

atribuir a ciertos modelos un anclaje referencial limitado por vías indirectas y selectivas. Sin embargo, este refuerzo no implica de forma necesaria una comprensión robusta, porque que haya referencia o corrección bajo criterios externos no equivale a que el sistema se comporte como agente epistémico responsable, capaz de regular su discurso por razones propias, sostener compromisos y discriminar de manera fiable entre saber, conjeturar e improvisar.

En suma, el grounding, tal como lo articulan Coelho Mollo y Millière (2025), fortalece la idea de un nivel intermedio, suficiente para robustecer la aceptación semántica del funcionamiento del modelo, pero insuficiente para justificar la atribución de entendimiento pleno en sentido agencial y normativo. Con todo, admitir que puede haber un anclaje referencial indirecto no resuelve por sí solo la cuestión del significado literal. Aun si los LLMs están conectados al mundo por cadenas mediadas y criterios selectivos, todavía puede sostenerse que sus outputs carecen de aquello que fija qué se dijo en una ocasión de uso. Esa es precisamente la línea que radicaliza el argumento de la falta de significado.

6. Falta de Significado de los LLMs

La tesis del entendimiento superficial no solo debe explicar por qué los LLMs "parecen entender", sino también resistir la crítica de que ese parecido no alcanza siquiera el umbral del significado literal. En este punto, una objeción reciente vuelve más exigente el debate: en lugar de negar la agencia epistémica del modelo, cuestiona que haya contenido semántico propiamente dicho en sus emisiones. En un artículo de reciente publicación se presentó un contraargumento que podría ser especialmente fuerte contra la atribución de entendimiento superficial y que me gustaría atender, el cual sostiene que el éxito lingüístico de los LLMs no es, en sentido estricto, éxito semántico: las salidas de estos sistemas serían fundamentalmente carentes de significado literal. Hattangadi y Schoubye (2025) formulan esta conclusión mediante un argumento

bastante simple basado en 2 premisas: la primera indica que, en usos concretos del lenguaje natural, ciertas intenciones comunicativas, y las actitudes que asociamos son necesarias para fijar qué se dijo literalmente; y segundo, que los LLMs no pueden plausiblemente poseer las clases relevantes de intenciones. De ahí infieren que las respuestas de los LLMs no tienen significado en el sentido literal que nos interesa cuando evaluamos qué se afirma, qué es verdadero o falso, y qué contenido proposicional queda determinado en una ocasión de uso.

La motivación de la primera premisa es la ubicuidad de lo que los autores llaman incertidumbre interpretativa, esta se refiere a que no se trata solo de casos exóticos, sino de fenómenos estructurales del lenguaje ordinario que suelen incluir ambigüedad léxica, ambigüedad estructural, anáfora, sensibilidad al contexto. En este tipo de casos, hay múltiples candidatos de contenido literal compatibles con la forma lingüística, y nuestra práctica ordinaria trata como decisivo el hecho de que el hablante pretendía decir una cosa y no otra.

A partir de ahí, Hattangadi y Schoubye sostienen que la segunda premisa no se apoya únicamente en una intuición searleana, sino en una tesis sobre arquitectura y entrenamiento, los LLMs carecerían de actitudes implícitas del tipo que permitiría fijar significado en la ocasión de emisión. Su argumento enfatiza que, durante el procesamiento, el texto es tokenizado y convertido en identificadores y vectores; en ese paso, información relevante para la interpretación semántica como por ejemplo, cuáles son unidades mínimas de significado se pierde, y además las mismas codificaciones vectoriales subdeterminan la desambiguación y la resolución de incertidumbre interpretativa, por lo que, aun si admitimos estados internos causales en los LLMs, esos estados no portarían el tipo de contenido que haría de soporte para intenciones referenciales o comunicativas, ni siquiera en un sentido ligero o implícito.

Lo que considero relevante de la posición de Hattangadi y Schoubye es que este enfoque intenta atacar, a la vez, dos rutas que se podrían tomar en de defensa del entendimiento superficial, por un lado, discuten la respuesta externalista según la cual bastaría la deferencia a la comunidad lingüística para fijar referencia y significado, y alegan que incluso la deferencia requiere alguna forma de intención, aunque sea ligera e implícita para tener consistencia con las convenciones correctas en la ocasión de uso. Por otro lado, discuten la respuesta internalista que apela a rol conceptual, que es justo el tipo de apoyo que Piantadosi y Hill presentan, y argumentan que la conformidad estructural con esquemas inferenciales no basta, porque esos esquemas son compatibles con múltiples interpretaciones no semánticas; sin una fijación proposicional previa, el rol no determina unívocamente significado, en consecuencia, las regularidades internas que motivan hablar de estructura conceptual parcial serían compatibles con que no haya hecho del asunto sobre qué se dijo en sentido literal.

La objeción de Hattangadi y Schoubye se basa en el argumento de que los esquemas inferenciales como ^I y ^E son plantillas sintácticas que pueden satisfacerse por operaciones no semánticas (como una compuerta AND en un circuito eléctrico); por tanto, que un sistema satisfaga esos patrones no prueba que sus estados porten contenido proposicional (Hattangadi & Schoubye, 2025). Sin embargo, este argumento se dirige contra una versión fuerte de CRS en la que el rol conceptual agota el significado. La versión moderada que adoptan Piantadosi y Hill —y que este artículo emplea— no afirma que la conformidad con esquemas inferenciales baste para significado pleno, sino que constituye un aspecto del significado. La pregunta, entonces, no es si la geometría vectorial del LLM es condición suficiente de contenido, ya que no lo es, sino, si exhibe una estructura que instancia parcialmente roles conceptuales genuinos. La evidencia empírica de isomorfismos representacionales entre vectores de LLMs y datos cognitivos humanos sugiere que sí, sin que ello resuelva la cuestión de si hay significado pleno o solo estructura parcial compatible con significado.

Finalmente, los autores explican por qué, a pesar de todo, los outputs parecen significativos y pueden ser epistémicamente útiles, puesto que lo que ocurre es una atribución por parte de los usuarios, análoga a leer como inglés una traza accidental. Interpretamos el texto como si hubiese sido producido por un hablante con las actitudes pertinentes, y al hacerlo proyectamos contenido y resolvemos ambigüedades usando claves contextuales, expectativas de cooperación y nuestros propios fines epistémicos. Esa utilidad no probaría significado literal del output, sino un tipo de significado sustituto, es decir un significado atribuido por nosotros mediante una suerte de pretensión interpretativa que maximiza beneficios epistémicos, aun cuando no haya sido el caso, desde el lado del sistema, sobre qué significó exactamente lo emitido.

Si este diagnóstico es correcto, entonces la discusión se desplaza: ya no se trata solo de qué hay "del lado del modelo", sino de cómo y con qué criterios los intérpretes estabilizan contenidos y evalúan desempeño. En particular, si el significado que atribuimos depende de prácticas interpretativas, se vuelve relevante preguntar quién está en posición de distinguir entre una imitación exitosa y una competencia genuina. Para dar sentido a esa intuición, conviene introducir una distinción sobre experticia y juicio competente.

7. Experticia, Juicio Competente en la Imitación

La respuesta al argumento de Hattangadi y Schoubye tiene dos dimensiones: la primera, desarrollada al final de la sección anterior, es semántica, se muestra que el argumento anti-CRS se dirige contra una versión fuerte de la semántica de rol conceptual que este artículo no adopta; la segunda, que desarrollo a continuación, es epistemológico-social: se refiere a las condiciones bajo las cuales los intérpretes atribuyen significado a los outputs y evalúan su éxito imitativo. El argumento de Hattangadi y Schoubye (2025) acierta al recordar que una parte sustantiva de lo que llamamos significado literal depende, en nuestras prácticas ordinarias, de la resolución de incertidumbres interpretativas

mediante actitudes e intenciones del hablante; por eso, cuando tales actitudes no están disponibles del lado del sistema, la tentación es concluir que el contenido del output es, en el mejor de los casos, una atribución del usuario. Sin embargo, esta conclusión negativa puede ser modulada si se toma en serio un hecho metodológico que el propio diagnóstico sugiere, cuando el contenido parece venir puesto por el intérprete, lo que está en juego no es solo la ontología del significado, sino la competencia del juez que lo atribuye y evalúa. Es aquí donde el marco de Collins y Evans (2007) resulta pertinente, porque permite describir con precisión por qué la eficacia imitativa del LLM no funciona como evidencia directa de entendimiento, sin por ello tener que degradar el output a sin sentido total.

Collins y Evans distinguen entre experticia contributiva y experticia interaccional, la primera es la capacidad de hacer dentro de un dominio, hacer física, construir el objeto original, intervenir competentemente en la práctica; la segunda es la capacidad de hablar el lenguaje del dominio con fluidez, reconociendo sus signos externos y participando en su conversación sin necesariamente poder contribuir en la práctica misma. Esta distinción se apoya en una estratificación más amplia del conocimiento que va desde el beer-mat knowledge hasta formas de comprensión más profundas, se puede repetir fórmulas que suenan explicativas sin que ello habilite a producir, corregir, decidir o intervenir de manera competente en el campo correspondiente (Collins & Evans, 2007). El punto crucial para el debate sobre LLMs es que la fluidez conversacional puede simular, ante ciertos públicos, la competencia real, y que la diferencia entre ambos niveles se vuelve visible solo cuando el juicio está en manos de quien posee la práctica y el trasfondo tácito del dominio.

Aplicado al contraargumento de Hattangadi y Schoubye, esto sugiere un replanteamiento, si el output parece significativo porque el usuario rellena intenciones y desambiguaciones, ello ocurre de manera especialmente marcada cuando el usuario opera con criterios de baja resolución, propios de una experticia meramente

interaccional. En esos casos, el LLM puede pasar por competente porque domina con notable éxito los marcadores superficiales del discurso, la jerga, las estructuras explicativas y los patrones retóricos que sostienen la impresión de contenido. Pero, justamente por eso, el éxito del modelo en convencer al usuario promedio no es una prueba de entendimiento; es, con frecuencia, una prueba de que el juez no tiene las herramientas para detectar la diferencia entre hablar como miembro de una práctica y participar realmente en ella. La imitación, entonces, se vuelve perfecta solo relativamente al nivel de experticia del evaluador.

Este marco refuerza la caracterización de los LLMs como fábricas de imitaciones lingüísticas, lo que el sistema produce son outputs con éxito interaccional, capaces de circular como habla humana ante un amplio rango de jueces, pero cuyo pasar por imitación depende de la asimetría entre experticia contributiva y experticia interaccional. Incluso si concedemos a Hattangadi y Schoubye que gran parte del contenido asignado al output es una atribución del intérprete, Collins y Evans permiten explicar por qué esa atribución es sistemáticamente más fácil en manos de quienes no están insertos en la práctica contributiva que estabiliza criterios de corrección, pertinencia y profundidad. En campos disputados, además, Collins y Evans subrayan que la distancia produce una certeza artificial, el público sin acceso al núcleo práctico del dominio tiende a evaluar con exceso de confianza, precisamente porque carece de contacto con las incertidumbres, los desacuerdos y los controles tácitos que organizan la producción de conocimiento.

En consecuencia primero, asumo que el argumento de la falta de significado deja de operar como un veredicto global y pasa a funcionar como advertencia metodológica, el output puede ser interpretado y usado como significativo, pero su evaluación epistémica depende de quién juzga y de qué criterios se activan; segundo, este diagnóstico encaja con la noción de entendimiento superficial, ya que el LLM puede sostener un desempeño lingüístico altamente competente en el plano interaccional y, en ese sentido,

producir textos con significado derivado en prácticas humanas, sin por ello adquirir la experticia contributiva asociada a participación en formas de vida, control normativo robusto y responsabilidad epistémica. Así, frente a Hattangadi y Schoubye, la respuesta no consiste en negar la dimensión proyectiva del usuario, sino en ubicarla donde corresponde: como parte de un régimen social de interpretación en el que la imitabilidad del output y la detectabilidad del engaño están moduladas por la estructura de la experticia. El resultado no es que "todo sea vacío", sino que el parecido a comprensión es un fenómeno real, pero competencia-relativo y, por ello, compatible con la tesis de un entendimiento superficial antes que robusto.

8. Entendimiento Superficial

Propongo llamar entendimiento superficial a un tipo de éxito lingüístico que, sin ser meramente vacío, tampoco alcanza el perfil que justificaría atribuir comprensión robusta al sistema. La noción pretende capturar un punto intermedio adecuado para LLMs: el modelo produce textos que funcionan semántica y pragmáticamente para lectores humanos, y además exhibe regularidades internas que sostienen asociaciones e inferencias; pero carece de rasgos normativos y epistémicos que caracterizan al entendimiento pleno. La ventaja de esta categoría es que permite describir con precisión la mezcla real que observamos en estos sistemas: resultados lingüísticos convincentes junto con fallas sistemáticas cuando exigimos trazabilidad, control de verdad o responsabilidad.

En ese sentido, el entendimiento superficial se alinea con la caracterización de Boisseau, según la cual el LLM se entiende mejor como una fábrica de productos lingüísticos con estatus de imitación del habla humana, más que como un agente que imita en sentido conductual (Boisseau, 2024). Ninguno de los autores previamente discutidos propone la categoría de entendimiento superficial directamente, Boisseau se posiciona como defensora del

no-entendimiento, Borg defiende el significado del producto, pero niega la aserción y deja abierto el estatus representacional interno, Piantadosi y Hill argumentan a favor de significado parcial, no de entendimiento.

La categoría que propongo es, por tanto, una síntesis propia que se justifica del siguiente modo: si el producto tiene contenido semántico derivado (Borg) y el proceso exhibe estructura de rol conceptual empíricamente sustentada (Piantadosi y Hill) y en ciertos casos hay anclaje referencial indirecto (Coelho Mollo y Millière), entonces atribuir al sistema un entendimiento nulo sería tan impreciso como atribuirle comprensión robusta. La categoría de entendimiento superficial nombra precisamente ese territorio intermedio: hay más que mera manipulación sintáctica ciega, pero menos que comprensión agencial. Por ello propongo las siguientes condiciones para atribuir entendimiento superficial a un sistema S respecto de un dominio discursivo D:

Condiciones positivas conjuntamente necesarias:

(C1) Suficiencia semántico-pragmática del producto: los outputs de S en D son interpretables por hablantes competentes, mantienen coherencia discursiva y realizan con éxito funciones comunicativas estándar. El contenido de estos outputs es semánticamente evaluable en virtud de deferencia semántica a nivel de tipo lingüístico (Borg, 2025).

(C2) Estructura interna parcialmente semántica: los estados internos de S exhiben patrones estables de asociación e inferencia que instancian, al menos parcialmente, roles conceptuales. La evidencia de esta condición proviene de isomorfismos entre la geometría representacional de S y la estructura conceptual documentada por medidas cognitivas y lingüísticas (Piantadosi & Hill, 2022).

(C3) Anclaje referencial limitado (condición gradual): algunos estados internos de S mantienen conexiones causales-informacionales indirectas con regularidades del mundo, mediadas por los datos textuales de entrenamiento, y han sido seleccionados por criterios que favorecen la preservación de información correcta (Coelho Mollo & Millière, 2025).

Condición negativa que restringe la atribución solo a superficial:

(C4) Carencia de normatividad agencial: S no regula su producción por compromisos epistémicos propios, no discrimina de manera fiable entre saber, conjeturar e improvisar, no asume responsabilidad por el contenido, y su orientación operativa es la predicción estadística, no la verdad.

Generando la siguiente definición: S exhibe entendimiento superficial respecto de D si y solo si satisface C1, C2 y en grado variable C3, mientras que C4 permanece vigente —es decir, el sistema carece de normatividad agencial.

Las condiciones positivas tienen apoyo diferenciado en la literatura revisada: C1 se apoya en el argumento de Borg (2025) sobre significado derivado a nivel de tipo lingüístico; C2 se apoya en la evidencia empírica y en el marco teórico de CRS que presentan Piantadosi y Hill (2022); C3 se apoya en el argumento sobre grounding referencial de Coelho Mollo y Millière (2025). Mientras que la condición negativa C4 recoge lo que ninguno de los tres apoyos resuelve: la carencia de gobernanza normativa y epistémica del sistema.

Que el modelo pueda generar afirmaciones con alta fluidez incluso cuando no dispone de base para ellas, o mantener la forma de una justificación sin controlar su validez, no demuestra que no haya significado en el output, pero sí que falta el tipo de control agencial que convertiría la producción lingüística en una práctica

epistémicamente responsable. Además, esta limitación tiene una dimensión metodológica: el "pasar por" competente depende de la competencia del evaluador, pues la fluidez interaccional puede ser suficiente para engañar a quienes carecen de experticia contributiva en el dominio relevante (Collins & Evans, 2007).

La expresión *imitación de significado* sintetiza esta situación, el modelo genera textos que parecen portar significado pleno porque se apoyan en la semanticidad derivada del lenguaje humano (Borg, 2025), en la estructura de rol conceptual parcial del proceso (Piantadosi & Hill, 2022) y, en ciertos casos, en formas indirectas de grounding (Coelho Mollo & Millière, 2025). Pero esa apariencia de plenitud semántica se sostiene sin los rasgos que normalmente anclan el significado en prácticas de agencia: intención comunicativa propia, responsabilidad por el contenido, control epistémico y capacidad de revisión por razones.

Frente al desafío de Hattangadi y Schoubye, la conclusión no necesita ser la nulidad semántica total; basta con afirmar que el rendimiento del modelo sostiene un nivel intermedio de éxito interpretativo y estructural, mientras la fijación fina del contenido literal y su evaluación epistémica permanecen, en parte, dependientes de prácticas humanas y de jueces competentes.

9. Entendimiento Superficial Frente al Bibliotechnism

La tesis defendida en este artículo guarda afinidades con la posición de Lederman y Mahowald (2024) bajo el nombre de bibliotechnism: la tesis de que los LLMs son tecnologías culturales —análogas a fotocopiadoras o prensas— que transmiten información sin crear contenido nuevo y sin poseer creencias, deseos ni intenciones. El bibliotechnism comparte con el marco de Boisseau (2024) la negación de agencia al sistema y la localización del significado en la relación entre el output y las prácticas humanas que le dan sentido. En varios aspectos, la posición de este artículo puede entenderse como una variante del bibliotechnism: sostengo,

con Lederman y Mahowald, que los outputs de los LLMs son significativos de manera derivada, no básica, y que la atribución de actitudes propositivas al sistema no es necesaria para explicar la mayor parte de su rendimiento lingüístico.

Lederman y Mahowald (2024) precisan el mecanismo por el cual los outputs heredan significado, y su propuesta complementa la noción de deferencia semántica de Borg (2025) empleada en la sección 4. Según ellos, un token producido por un LLM es significativo derivativamente si existe una cadena causal apropiada que lo conecta con tokens originales en los datos de entrenamiento producidos por agentes humanos que si tendrían significado básico, y si el proceso de generación es causalmente sensible a la inteligibilidad de los datos. En su sentido técnico, un token es inteligible si y solo si es posible que alguien lo entienda según las convenciones del idioma. El test contrafáctico que proponen es revelador: si el LLM hubiera sido entrenado con galimatías, produciría galimatías; si se entrena con texto inteligible, produce texto inteligible. Esta sensibilidad contrafáctica distingue a los LLMs de modelos más simples, como los unigramas, cuyo output no preserva la cohesión que hace significativas a las expresiones complejas.

Este argumento refuerza la condición C1 del entendimiento superficial referido a la suficiencia semántico-pragmática del producto, al ofrecer un mecanismo causal-histórico preciso para explicar por qué los outputs tienen contenido derivado incluso cuando son textos novedosos, es decir, textos que no aparecen literalmente en los datos de entrenamiento. Al mismo tiempo, el argumento proporciona una respuesta adicional a la analogía de Hattangadi y Schoubye (2025), quienes comparan los outputs de los LLMs con la traza accidental de una hormiga en la arena. La diferencia relevante entre la traza de la hormiga y el token del LLM es precisamente la cadena causal apropiada y la sensibilidad a la inteligibilidad: la primera carece de ambas, el segundo las satisface.

Sin embargo, el bibliotechnism enfrenta un desafío que Lederman y Mahowald formulan con claridad y que la tesis de entendimiento superficial reconoce como límite, los LLMs pueden generar nombres nuevos que parecen referir a entidades inéditas: por ejemplo, inventar un nombre ficticio y usarlo consistentemente para describir hechos de un personaje histórico, o crear un diagrama y nombrar coherentemente sus elementos. En estos casos, la referencia de los nuevos tokens no puede derivarse de tokens en los datos de entrenamiento, porque por hipótesis no existen tokens allí que refieran a esas entidades con esos nombres. Este es el problema de la referencia novedosa o nobel reference.

Las condiciones C1–C3 del entendimiento superficial no resuelven plenamente este problema. C1, referida al significado derivado, depende de cadenas causales con los datos, y el novel reference rompe esas cadenas para los nombres nuevos. C2, referida a la estructura de rol conceptual, podría ofrecer una respuesta parcial: si el modelo opera sobre una red de relaciones conceptuales, la introducción de un nombre nuevo para una configuración particular puede entenderse como una operación sobre esa estructura relacional, sin necesidad de postular intenciones referenciales, pero esta respuesta es tentativa y requiere más desarrollo.

Lederman y Mahowald proponen que el interpretationism (Dennett, 1971; Davidson, 1973) ofrece una salida al problema de la referencia novedosa: si el comportamiento de un sistema se explica bien por la hipótesis de que tiene creencias, deseos e intenciones, entonces el sistema las tiene. Bajo esta tesis, los LLMs podrían poseer actitudes en un sentido funcional ligero que no requiere consciencia ni sentencia, y el novel reference se resolvería aplicando las mismas teorías de introducción de nombres que se aplican a los humanos.

El interpretacionismo constituye una alternativa genuina a la posición de este artículo, ya que se lo adopta, la categoría de entendimiento superficial se vuelve innecesaria: el sistema directamente tendría actitudes, aunque solo ligeras, y la cuestión del significado del output se resolvería por vía estándar. Sin embargo, opté por no adoptar esta ruta por dos razones: primera, el interpretacionismo es una posición sustantiva y controvertida en filosofía de la mente (Schwitzgebel, 2023), y basar el argumento en ella convertiría la tesis sobre LLMs en una tesis sobre la naturaleza de las actitudes propositivas, desplazando el foco del artículo; segunda, la tesis de entendimiento superficial pretende ser neutral respecto de si el interpretacionismo es correcto: busca describir lo que podemos afirmar sobre los LLMs antes de resolver ese debate metafísico. Es decir, el entendimiento superficial se concibe como una categoría que resulta útil mientras tanto, compatible con que el interpretacionismo termine por atribuir actitudes a los LLMs, pero también compatible con que no lo haga.

En consecuencia, el problema de la referencia novedosa queda reconocido como un límite abierto de la posición aquí defendida, la tesis de entendimiento superficial cubre la mayor parte del rendimiento lingüístico de los LLMs, como el ensamblaje novedoso de significados derivados, regularidades inferenciales internas, anclaje referencial indirecto, pero no ofrece todavía una explicación completa de los casos en que el modelo parece crear referencia genuinamente nueva. Este es un problema que comparten, en distintos grados, todas las posiciones del debate que no atribuyen actitudes a los LLMs.

Conclusiones

El hallazgo central es que la discusión sobre si los LLMs entienden se vuelve más clara cuando se separan tres niveles de análisis: el estatus semántico del producto, la estructura del proceso y la agencia del sistema. Con la distinción de Boisseau, se establece que el producto tiene estatus de imitación y no implica por sí solo comprensión. Con Borg, se justifica que el producto porte contenido semántico derivado a nivel de tipo lingüístico. Con Piantadosi y Hill, se vuelve plausible que el proceso instancie parcialmente roles conceptuales, según evidencia de isomorfismos representacionales. Con Coelho Mollo y Millière, se admite un grado variable de anclaje referencial indirecto. Ninguna de estas líneas propone por sí sola la categoría de entendimiento superficial: esta es una síntesis propia de este artículo, que define un territorio intermedio entre la manipulación sintáctica ciega y la comprensión robusta. Frente a la tesis de Hattangadi y Schoubye, se ofrece una doble respuesta: semántica, al mostrar que su argumento anti-CRS no alcanza a la versión moderada aquí adoptada; y epistemológica, al mostrar con Collins y Evans que la evaluación del éxito imitativo depende de la competencia del juez.

La categoría de entendimiento superficial, definida por las condiciones C1–C4, permite explicar el éxito semántico-pragmático y las regularidades inferenciales de los LLMs, sin atribuir responsabilidad, control de verdad ni comprensión agencial. Finalmente, la confrontación con el bibliotechnism de Lederman y Mahowald (2024) confirma la viabilidad de la posición intermedia, al tiempo que señala su principal límite pendiente en la referencia novedosa. El mecanismo de significado derivado por cadenas causales y sensibilidad a la inteligibilidad refuerza la explicación del contenido del producto; el problema de la referencia novedosa, en cambio, marca un punto donde la tesis de entendimiento superficial requiere todavía desarrollo adicional, posiblemente en diálogo con el interpretationism o con teorías de la referencia que no presupongan agencia que comprometerían un debate sobre las actitudes de los LLMs que excede los fines de este texto.

Referencias Bibliográficas

Block, N. (1986). Advertisement for a semantics for psychology. *Midwest Studies in Philosophy, 10*(1), 615–678.

Boisseau, É. (2024). Imitation and large language models. *Minds & Machines, 34*, Article 42. https://doi.org/10.1007/s11023-024-09698-6

Borg, E. (2004). *Minimal Semantics.* Oxford University Press.

Borg, E. (2025). LLMs, Turing tests and Chinese rooms: The prospects for meaning in large language models. *Inquiry*. Advance online publication. https://doi.org/10.1080/0020174X.2024.2446241

Collins, H. & Evans, R. (2007). *Rethinking Expertise.* University of Chicago Press. Coelho Mollo, D. & Millière, R. (2025). The vector grounding problem (arXiv:2304.01481v3). arXiv. https://doi.org/10.48550/arXiv.2304.01481

Davidson, D. (1973). Radical interpretation. *Dialectica, 27*, 313–328

Dennett, D. (1971). Intentional systems. *The Journal of Philosophy, 68*(4), 87–106.

Ferraris, A., Audrito, D., Di Caro, L., & Poncibò, C. (2025). The architecture of language: Understanding the mechanics behind LLMs. *Cambridge Forum on AI: Law and Governance, 1*, e11, 1–19. https://doi.org/10.1017/cfl.2024.16

Harman, G. (1987). (Non-solipsistic) conceptual role semantics. En E. Lepore (Ed.), *New Directions in Semantics* (pp. 55–81).

Harman, G. (1999). Conceptual role semantics. En G. Harman (Ed.), *Reasoning, meaning, and mind* (pp. 117–131). Oxford University Press.

Hattangadi, A. & Schoubye, A. (2025). The outputs of large language models are meaningless. En H. Cappelen & R. Sterken (Eds.), *Communicating with AI: Philosophical perspectives.* Oxford University Press.

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

Piantadosi, S., & Hill, F. (2022). Meaning without reference in large language models. *arXiv*. https://doi.org/10.48550/arXiv.2208.02957

Schwitzgebel, E. (2023). Belief. En E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of Philosophy*. Stanford University.

Searle, J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences, 3*(3), 417–424. https://doi.org/10.1017/S0140525X00005756

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