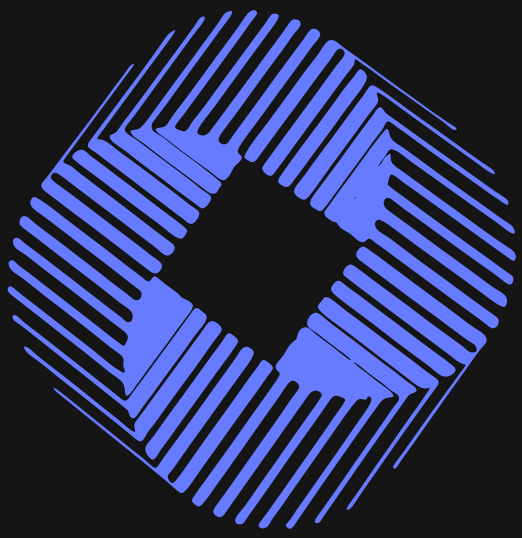
TecnoPsique.

LECTURAS / TRADUCCIONES



Seis ideas equivocadas sobre los grandes modelos de lenguaje

Un modelo mínimo y una taxonomía diagnóstica

Zhicheng Lin

Traducción íntegra al español · PNAS Nexus, 2026

Seis ideas equivocadas sobre los grandes modelos de lenguaje

Un modelo mínimo y una taxonomía diagnóstica

Zhicheng Lin

Departamento de Psicología, Universidad Yonsei, Seúl 03722, República de Corea. Correspondencia: zhichenglin@gmail.com. Editor del artículo original: Derek Abbott.

Artículo de perspectiva publicado en PNAS Nexus, 2026, 5(7), pgag236. Publicación: 8 de julio de 2026. Recibido: 20 de diciembre de 2025. Aceptado: 29 de junio de 2026.

Sobre esta edición — nota de TecnoPsique

Traducción íntegra al español, no oficial, realizada con asistencia de IA para TecnoPsique. El texto y sus argumentos pertenecen a Zhicheng Lin; esta edición no implica su revisión ni su respaldo a la traducción. Se tradujeron el cuerpo del artículo, las tablas, los recuadros y los pies de figura. Las tablas se redistribuyeron en fichas para facilitar la lectura. Se conserva la figura original con una clave de lectura en español; las referencias mantienen sus títulos y datos bibliográficos en el idioma original. Se añadieron portada, esta nota y cambios de diagramación. No se añadieron resultados ni argumentos al artículo.

Referencia del original: Lin, Z. (2026). Six misconceptions about large language models: A minimal model and diagnostic taxonomy. PNAS Nexus, 5(7), pgag236. https://doi.org/10.1093/pnasnexus/pgag236

Criterio terminológico: se conserva la sigla inglesa LLM para «gran modelo de lenguaje» y RLHF para «aprendizaje por refuerzo a partir de retroalimentación humana». «Token» designa una unidad de la secuencia procesada por el modelo. Las citas entre comillas en el cuerpo y en las tablas también están traducidas. Los números entre paréntesis remiten a la bibliografía original.

Resumen

Los grandes modelos de lenguaje (LLM) ya están integrados en los flujos de trabajo científicos, educativos y de gobernanza, y los debates se concentran en sus capacidades, mecanismos e impactos. Sin embargo, esos debates siguen estructurados por persistentes teorías intuitivas: modelos explicativos informales que orientan actitudes y acciones. Tanto las consignas deflacionarias («solo autocompletado», «loros estocásticos» y «el promedio de internet») como los encuadres antropomórficos («agentes emergentes» y «protomentes») captan rasgos genuinos de los sistemas actuales, pero confunden esos rasgos con el conjunto. Este artículo de perspectiva propone un modelo mínimo de trabajo de los sistemas basados en LLM, centrado en cuatro distinciones: entre el preentrenamiento y los sistemas desplegados; entre la distribución aprendida y las muestras particulares; entre la memoria paramétrica, la contextual y la externa; y entre la competencia para realizar tareas y la agencia. El modelo se utiliza para diagnosticar seis ideas equivocadas sobre los LLM: la predicción del siguiente token, la regresión a la media, la reproducción de datos de entrenamiento, la memoria del modelo, el alineamiento y la comprensión. Para cada una, el análisis identifica qué tiene de acertado, qué distinciones confunde y qué consecuencias se desprenden para la evaluación de capacidades, el diseño de sistemas y la gobernanza. Aplicado a las políticas editoriales sobre IA como casos de estudio de gobernanza, el marco muestra cómo su lenguaje puede confundir estas distinciones y cómo corregir esos errores. Al tratar a los LLM como simuladores del discurso y del desempeño en tareas, el modelo evita así la oposición binaria entre loro y mente, y ofrece un conjunto de herramientas diagnósticas para localizar y corregir los errores que estas teorías intuitivas perpetúan.

Palabras clave: grandes modelos de lenguaje; ideas equivocadas; despliegue; alineamiento; memoria.

Teorías intuitivas sobre los grandes modelos de lenguaje

Hoy son pocos los debates sobre la práctica científica, la educación o el trabajo creativo que no invocan a los grandes modelos de lenguaje (LLM). Periodistas, críticos e investigadores recurren a consignas prefabricadas para declarar qué son «realmente» estos sistemas: «autocompletado glorificado» o «simples predictores del siguiente token»; «loros estocásticos»; «algoritmos de compresión de texto con pérdida» o «un JPEG borroso de todo el texto de la Web»; «generadores de palabrería»; «trucos de salón de alta tecnología»; o «el promedio de internet, retocado en su tono» (1–3). Otros los presentan en términos más antropomórficos: como «razonadores sobrehumanos», «protoagentes» o manifestaciones de la «sabiduría de la multitud de silicio» (4–7).

Estas consignas pueden transformarse en teorías intuitivas: modelos explicativos informales —a menudo parciales y tácitos— utilizados para dar sentido a los sistemas tecnológicos y orientar las acciones hacia ellos (8, 9). «Intuitivas» se refiere al modo de explicación, no al grado de sofisticación de quien habla; de hecho, las explicaciones informales y las técnicas pueden coexistir en investigadores, responsables de políticas y usuarios legos. Varias de las

expresiones anteriores también comenzaron como analogías o argumentos de alcance delimitado: «loros estocásticos» sintetizaba una amplia crítica sociotécnica de la escala, la procedencia de los datos y los daños del texto sintético de apariencia humana, en la cual la afirmación de que los modelos ensamblan formas lingüísticas sin referencia al significado era solo una parte (1); «un JPEG borroso de la Web» era una analogía de compresión (3); y «generadores de palabrería» apelaba al sentido técnico de indiferencia hacia la verdad propuesto por Frankfurt (2). Se convierten en teorías intuitivas cuando se separan de su alcance original y se reutilizan como explicaciones generales: se aísla un rasgo específico de los sistemas actuales —el entrenamiento basado en datos, los objetivos de predicción del siguiente token, la falta de corporalidad o la ausencia de aprendizaje persistente en línea— y luego se lo amplifica hasta convertirlo en una explicación global de lo que los sistemas basados en LLM son o podrían llegar a ser. El resultado es una polarización peculiar: las posiciones deflacionarias descartan a los LLM como triviales o los presentan como potentes simuladores carentes de comprensión, mientras que las posiciones antropomórficas los tratan como mentes o agentes incipientes. Ambos extremos pueden desorientar la evaluación de capacidades, el diseño de sistemas y la gobernanza.

Estas teorías intuitivas reaparecen en la publicación científica, los medios de comunicación, el razonamiento de los usuarios y las orientaciones institucionales (10). El lenguaje antropomórfico en los artículos de investigación en informática ha aumentado con el tiempo y se amplifica en la cobertura periodística posterior (11), con un giro, tras ChatGPT, hacia encuadres centrados en el peligro y la antropomorfización (12, 13). Estudios representativos a escala nacional muestran que las metáforas públicas sobre la IA tienen una estructura, son medibles e inciden en la confianza y la adopción, y que la semejanza humana percibida aumenta con el tiempo (14). Los estudios con usuarios muestran distorsiones paralelas en los modelos mentales sobre cómo los sistemas basados en LLM generan, recuperan y almacenan información: los usuarios confunden la generación con la búsqueda (15), malinterpretan lo que almacenan y reutilizan las diferentes capas de memoria (16), y modifican sus juicios sobre precisión y riesgo en respuesta a señales antropomórficas de la interfaz (17, 18). El trabajo con grupos focales reconstruye las teorías intuitivas de personas legas sobre los chatbots de IA generativa y muestra cómo esas teorías configuran sus estrategias de interacción (19). El mismo patrón aparece en las orientaciones institucionales: los análisis de la guía de la UNESCO sobre IA generativa encuentran metáforas de personificación incorporadas en su explicación del comportamiento de los modelos (20).

El patrón es inequívoco, pero su estructura conceptual no ha sido delimitada sistemáticamente. Las revisiones generales de los modelos fundacionales catalogan capacidades, aplicaciones y riesgos sociales, pero no rastrean las confusiones conceptuales hasta los errores que producen (21). Las taxonomías de riesgos clasifican los daños por tipo y gravedad sin diagnosticar las teorías intuitivas que configuran el razonamiento sobre ellos (22). Los trabajos sobre antropomorfismo en IA e interacción persona-computadora identifican los riesgos de atribuir estados mentales a los sistemas, pero concentran esos riesgos en las capacidades comunicativas —persuasión, empatía y representación de roles— más que en la arquitectura de memoria, el alineamiento o el lenguaje de las políticas

institucionales (23, 24). Las revisiones sobre memoria y generación aumentada por recuperación ofrecen taxonomías técnicas detalladas de los mecanismos de almacenamiento y recuperación, pero no conectan las ideas equivocadas sobre esos mecanismos con los errores posteriores en la evaluación de capacidades, el diseño del despliegue o el razonamiento sobre políticas (25, 26).

Este artículo de perspectiva desarrolla una taxonomía diagnóstica de seis ideas equivocadas presentes en el discurso científico, público e institucional, incluidas las políticas editoriales actuales sobre IA generativa. La taxonomía se organiza en torno a cuatro distinciones: entre el preentrenamiento y los sistemas desplegados; entre la distribución aprendida y las muestras particulares; entre la memoria paramétrica, la contextual y la externa; y entre la competencia para realizar tareas y la agencia. Aplicado como matriz diagnóstica (Tabla 1), este marco muestra cómo las confusiones conceptuales producen errores previsibles en la evaluación de capacidades, el diseño del despliegue y las políticas institucionales sobre IA (Tabla 2).

Un modelo mínimo de trabajo de los sistemas basados en LLM

Diagnosticar estos malentendidos requiere un modelo mínimo de trabajo sobre cómo se entrenan y despliegan los LLM. En todo el texto, «LLM» se refiere a la clase de modelos; «sistema basado en LLM», a un producto desplegado que se construye alrededor de uno de esos modelos; e «IA» e «IA generativa», al campo más amplio, al lenguaje de las políticas citadas o a casos de comparación que no son LLM. Los componentes básicos de los LLM están bien documentados en la literatura técnica (27–37); la Figura 1 los organiza en torno a cuatro distinciones que hacen explícita la estructura conceptual e identifican dónde suelen omitirse distinciones clave.

Durante el preentrenamiento, un LLM recibe una secuencia de tokens y devuelve una distribución sobre el siguiente token (27). El objetivo estándar minimiza la entropía cruzada del siguiente token sobre un corpus muy grande de texto y, cada vez más, de otras modalidades. El resultado no es una tabla de continuaciones prefabricadas, sino una aproximación de alta dimensionalidad a la distribución condicional del lenguaje dado un contexto. El preentrenamiento ocurre fuera de línea: una vez terminado el entrenamiento y fijados los pesos, el modelo base no sigue aprendiendo durante la inferencia o el despliegue (28). De aquí surge la primera distinción (A): entre el régimen de entrenamiento y el comportamiento de los sistemas desplegados construidos sobre el modelo entrenado.

En el momento de la inferencia, el modelo recibe una secuencia de entrada —instrucciones, historial del diálogo y cualquier contenido recuperado— y devuelve una distribución de probabilidad sobre los siguientes tokens. Luego se genera una secuencia específica mediante una política de decodificación —selección voraz, muestreo por temperatura, muestreo de núcleo o top-p, o sus variantes (29)—, cada una de las cuales selecciona elementos de la distribución aprendida o la modifica, estableciendo un compromiso entre diversidad y previsibilidad. De aquí surge la segunda distinción (B): entre la distribución que el modelo ha aprendido y las muestras particulares observadas bajo un régimen de decodificación elegido.

Configuraciones predeterminadas como la decodificación a baja temperatura y los prompts de sistema extensos y cautelosos crean la familiar «voz de asistente», pero ninguno de estos rasgos es intrínseco al modelo en sí.

Tampoco suele tratarse de la distribución de un modelo base sin ajustes. La mayoría de los modelos en producción son variantes ajustadas. El ajuste supervisado por instrucciones utiliza pares de instrucción y respuesta seleccionados; el aprendizaje por refuerzo a partir de retroalimentación humana (RLHF) utiliza señales de recompensa derivadas de preferencias: ambos actualizan directamente los parámetros del modelo (30). Estos procedimientos modifican la distribución condicional y orientan al modelo hacia respuestas que los evaluadores humanos consideran más útiles, inocuas u honestas (31). Estas actualizaciones no añaden un filtro desmontable a un núcleo neutral, sino que inscriben nuevas regularidades y patrones de evitación en los mismos pesos que codifican conocimientos y competencias (32). Por ello, distintas variantes ajustadas de un mismo modelo base pueden tener perfiles de comportamiento diferentes, aun compartiendo arquitectura e historia de preentrenamiento.

Cuando se integra en un producto basado en LLM, el modelo se convierte en un componente, aunque central. Un asistente de chat, una herramienta de programación o un entorno de «agentes» rodea al modelo de prompts de sistema, interfaces de usuario, filtros de seguridad, API de herramientas, sistemas de recuperación y, a veces, actuadores en entornos externos. El sistema puede interpretar las salidas del modelo como acciones —llamadas a herramientas, consultas a bases de datos y movimientos en un entorno— cuyos resultados se incorporan luego al siguiente prompt, creando un bucle cerrado (33, 34). Este bucle puede funcionar como un controlador de toma de decisiones, aunque cada llamada al modelo siga consistiendo en predecir el siguiente token (35). Por tanto, un mismo modelo base preentrenado puede sostener sistemas cualitativamente distintos según cómo se lo integre, alinee y despliegue.

Una tercera distinción (C) se refiere a cómo los sistemas basados en LLM retienen, recuperan y reutilizan información entre interacciones, y separa tres tipos de «memoria» (25, 26, 28, 36, 38). Primero, la memoria paramétrica: información codificada implícitamente en los pesos como una compresión con pérdida y sesgos de los datos de entrenamiento. Segundo, la memoria contextual: la secuencia de entrada actual, incluido el historial de conversación y el contenido recuperado, que funciona como una memoria de trabajo acotada dentro de la ventana de contexto. Tercero, la memoria externa, propia del producto: registros, perfiles de usuario, bases de conocimiento y almacenes vectoriales que se mantienen fuera del modelo y se incorporan selectivamente a los prompts mediante recuperación o código de orquestación. Las confusiones habituales sobre el «recuerdo», el «olvido» y el «aprendizaje a partir de chats» de los LLM surgen de confundir estas tres capas.

Por último, la cuarta distinción (D) concierne al estatus cognitivo. Los LLM son simuladores de gran capacidad del discurso y del desempeño en tareas: codifican representaciones internas ricas y permiten niveles importantes de abstracción, transferencia y aprendizaje en contexto entre distintos dominios (39). Al mismo tiempo, estos sistemas no poseen

comprensión de tipo humano, creencias o intenciones unificadas ni conciencia fenoménica (37). Esta distinción separa la comprensión funcional —aproximadamente, la capacidad de utilizar información de manera adecuada en distintas tareas— de una comprensión de tipo humano, corporizada y experiencial. Por eso, los términos de agencia —creencia, deseo e intención— deberían reservarse para funciones cuidadosamente especificadas, en lugar de emplearse como atribuciones informales (40, 41).

Este modelo mínimo funciona como una herramienta conceptual de análisis (42), conservando las distinciones críticas necesarias para localizar los errores que diagnostica (43). Los productos desplegados actuales suelen quedar dentro de su alcance porque su comportamiento está configurado por alguna combinación de los componentes que especifica: capas de integración, herramientas, recuperación, memoria externa, políticas de decodificación y acción en bucle cerrado. Una afirmación constituye, por tanto, una idea equivocada cuando omite una de estas distinciones e induce una inferencia errónea sobre el sistema en cuestión.

Tabla 1. Matriz diagnóstica de seis ideas equivocadas sobre los LLM

Contenido original redistribuido en fichas. Cada ficha conserva las seis columnas de la matriz.

Estadística y generatividad

1. Los LLM son solo autocompletado o loros estocásticos

Núcleo de verdad. Los modelos base se entrenan para predecir el siguiente token en grandes corpus.

Distinción confundida. A: objetivo de entrenamiento frente a sistema desplegado.

Error resultante. Tomar la predicción del siguiente token como una explicación cognitiva completa; descartar la competencia a nivel de sistema y el uso de herramientas.

Pregunta diagnóstica. ¿Cómo cambia el comportamiento cuando el modelo se integra con herramientas o en sistemas de bucle cerrado?

Corrección práctica. Especificar siempre la capa de integración del sistema, sus herramientas y la política de decodificación al describir sus capacidades.

Estadística y generatividad

2. Los LLM son el promedio de internet o regresan a la media

Núcleo de verdad. El preentrenamiento aprende una distribución de alta dimensionalidad sobre las continuaciones.

Distinción confundida. B: distribución frente a muestra.

Error resultante. Suponer que una salida insípida y homogeneizada es matemáticamente inevitable, en lugar de una decisión de producto.

Pregunta diagnóstica. ¿Qué ocurre si cambiamos los prompts, la temperatura y otros ajustes de muestreo?

Corrección práctica. Variar la decodificación y los prompts; informar qué regímenes utilizan las evaluaciones y los despliegues.

Estadística y generatividad

3. Los LLM solo memorizan los datos de entrenamiento

Núcleo de verdad. Los modelos son compresores generativos y sí presentan cierta reproducción literal.

Distinción confundida. C: memoria paramétrica frente a contextual y externa; compresión frente a recuperación.

Error resultante. Tratar todas las salidas como texto copiado; ignorar tanto la recombinación genuina como los riesgos específicos de memorización.

Pregunta diagnóstica. ¿Esta salida procede realmente de manera literal o casi literal del conjunto de entrenamiento?

Corrección práctica. Auditar la reproducción literal en dominios sensibles; seleccionar y cuidar los datos, y añadir medidas contra filtraciones y filtros cuando sea necesario.

Memoria y alineamiento

4. Los LLM recuerdan todo sobre mí frente a no recuerdan nada

Núcleo de verdad. Los pesos son fijos; la memoria de trabajo está limitada por la ventana de contexto; los productos pueden añadir almacenamiento separado.

Distinción confundida. C: memoria paramétrica frente a contextual y externa.

Error resultante. Confiar en exceso en el aparente aprendizaje a partir de chats o diseñar sistemas que suponen una ausencia total de estado.

Pregunta diagnóstica. ¿Dónde se almacena realmente la información en este producto y cómo se reutiliza?

Corrección práctica. Diseñar capas de memoria y políticas explícitas y auditables en cada nivel: pesos, contexto y almacenes externos.

Memoria y alineamiento

5. El RLHF o el ajuste fino son solo un filtro de seguridad desmontable sobre un núcleo neutral

Núcleo de verdad. El ajuste por instrucciones y el RLHF modifican los mismos pesos que codifican conocimientos y habilidades.

Distinción confundida. A: entrenamiento frente a despliegue; pesos del modelo frente a filtros externos.

Error resultante. Suponer que el alineamiento es reversible y no tiene costos; ignorar los valores incorporados y los compromisos entre capacidades.

Pregunta diagnóstica. ¿Qué comportamientos y métricas cambian cuando aplicamos o retiramos el ajuste fino, manteniendo fijos los filtros externos?

Corrección práctica. Tratar el ajuste fino como entrenamiento sustantivo: evaluar deterioros del rendimiento y cambios de valores; documentar sus objetivos por separado de los filtros usados durante el despliegue.

Estatus cognitivo

6. Los LLM piensan como humanos frente a carecen por completo de comprensión

Núcleo de verdad. Los modelos permiten una competencia considerable en tareas sin poseer mentes corporizadas de tipo humano.

Distinción confundida. D: competencia frente a agencia; comprensión funcional frente a comprensión de tipo humano.

Error resultante. Antropomorfizar los modelos como agentes con creencias y derechos o trivializar sus capacidades como mera repetición.

Pregunta diagnóstica. ¿Qué tareas concretas puede realizar fiablemente este sistema y qué formas de agencia le estamos atribuyendo implícitamente?

Corrección práctica. Describir las capacidades en términos funcionales; evitar tratar la fluidez estilística como prueba de creencias, deseos o estatus moral.

Cuatro distinciones centrales: A, objetivo de preentrenamiento frente a sistema desplegado; B, distribución aprendida frente a muestra particular; C, memoria paramétrica frente a contextual y externa; D, competencia en tareas frente a agencia. Cuando una idea equivocada confunde más de una distinción, se indica la principal. Las columnas avanzan desde la afirmación intuitiva, pasando por el diagnóstico —núcleo de verdad, distinción confundida y error resultante—, hasta el uso correctivo —pregunta diagnóstica y corrección práctica—.

Tabla 2. Aplicación de la matriz diagnóstica al lenguaje de las políticas editoriales sobre IA

Ejemplos de políticas al 1 de abril de 2026. Las citas se presentan traducidas al español.

1. Los LLM son solo autocompletado o loros estocásticos

Ejemplo de política. APS: «[Los LLM] pueden generar contenido lingüísticamente plausible, pero no científicamente plausible».

Distinción confundida. A: objetivo de entrenamiento frente a sistema desplegado. B: distribución frente a muestra.

Error resultante en el razonamiento de la política. Trata la orientación «lingüística» como intrínsecamente opuesta a lo «científicamente plausible», fomentando una visión de los LLM como meros hilvanadores de texto, en lugar de componentes de sistemas que pueden combinarse con herramientas, recuperación y verificación.

Pregunta diagnóstica y formulación mejorada. ¿Con qué prompts, decodificación y configuración de herramientas se produjo el contenido, y cómo se contrastó con las fuentes primarias?

2. Los LLM son el promedio de internet o regresan a la media

Ejemplo de política. SAGE: «Algunos LLM podrían haberse entrenado solo con datos hasta un año específico, lo que potencialmente daría lugar a un conocimiento incorrecto o incompleto de un tema»; APS: «Algunos LLM solo se entrenan con contenido publicado antes de una fecha determinada y, por tanto, presentan un panorama incompleto».

Distinción confundida. B: distribución frente a muestra. A: entrenamiento frente a sistema desplegado.

Error resultante en el razonamiento de la política. Implica que un «panorama incompleto» es una simple función de la fecha de corte del entrenamiento, en lugar de depender de cómo un sistema desplegado adquiere y valida efectivamente información para una tarea determinada.

Pregunta diagnóstica y formulación mejorada. ¿Este sistema específico depende solo del conocimiento paramétrico o utiliza recuperación de información actualizada? En ambos casos, ¿cómo se verifican las afirmaciones clave?

3. Los LLM solo memorizan los datos de entrenamiento

Ejemplo de política. SAGE: «Los LLM podrían reproducir inadvertidamente fragmentos significativos de texto de fuentes existentes sin la debida cita»; APS: «El LLM puede haber reproducido texto sustancial de otras fuentes».

Distinción confundida. C: memoria paramétrica frente a contextual y externa.

Error resultante en el razonamiento de la política. Centra el riesgo en la reproducción casi literal, reforzando una imagen de los LLM como tablas de consulta y restando énfasis a fallos más habituales, como la paráfrasis sin atribución o la copia estructural de argumentos y métodos.

Pregunta diagnóstica y formulación mejorada. ¿Esta salida es realmente casi literal o el riesgo más apremiante es una apropiación estructural o de ideas que también requiere citas y atribuciones explícitas?

4. Los LLM recuerdan todo sobre mí frente a no recuerdan nada

Ejemplo de política. APA: «La organización que opera la IA generativa probablemente tendrá acceso a [la información introducida]».

Distinción confundida. C: memoria paramétrica frente a contextual y externa. A: modelo frente a despliegue.

Error resultante en el razonamiento de la política. Equipara «introducir información en IA generativa» con otorgar amplio acceso a una organización monolítica, fundiendo las políticas de registro y retención a nivel de producto en un único riesgo indiferenciado asociado a la IA generativa como tal.

Pregunta diagnóstica y formulación mejorada. ¿Dónde se almacenan efectivamente estos datos —pesos, contexto, registros, bases de datos externas—, quién puede acceder a cada capa y bajo qué políticas de retención y reutilización para entrenamiento?

6. Los LLM piensan como humanos frente a carecen por completo de comprensión

Ejemplo de política. SAGE/APS: «[Los LLM y la IA generativa] no pueden reproducir el pensamiento creativo y crítico humano».

Distinción confundida. D: competencia frente a agencia.

Error resultante en el razonamiento de la política. Incorpora una tesis cognitiva fuerte al lenguaje de las políticas, al tratar el «pensamiento creativo y crítico humano» como una propiedad de todo o nada y fomentar una visión binaria en la que existe una comprensión plenamente humana o se descarta al sistema como no creativo.

Pregunta diagnóstica y formulación mejorada. ¿Qué tareas concretas o formas de desempeño «creativo» o «crítico» están en cuestión y qué fiabilidad tiene este sistema en ellas bajo condiciones de evaluación documentadas?

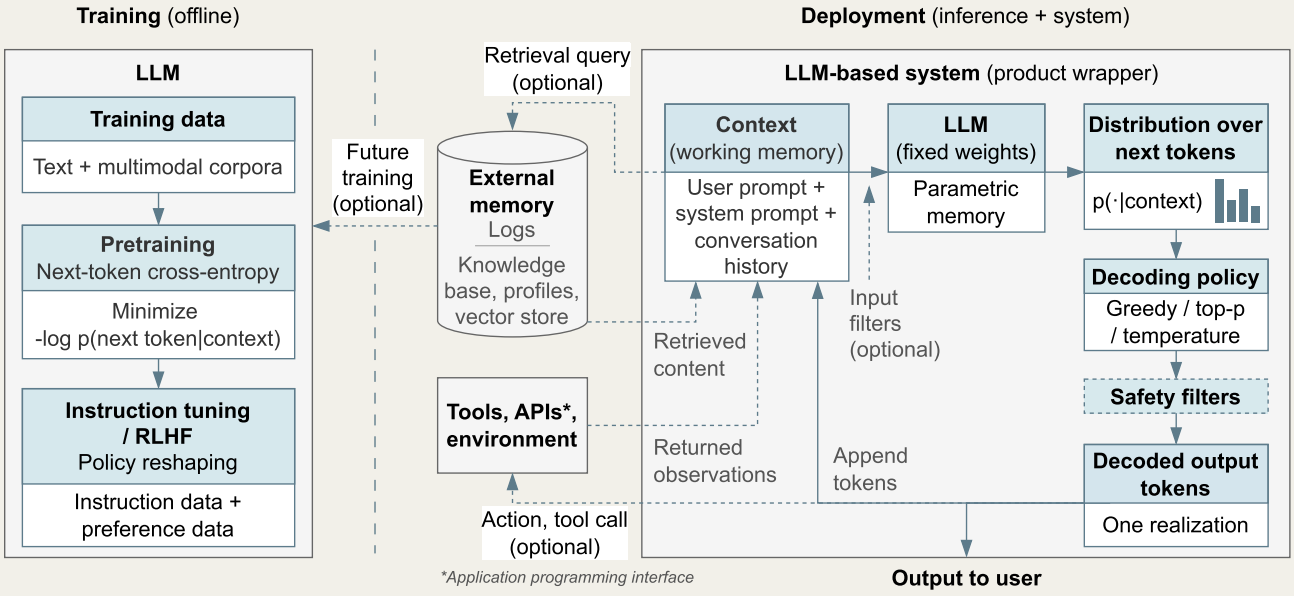
Cada ficha aplica el diagnóstico correspondiente de la Tabla 1 al lenguaje de las políticas editoriales. APS: Association for Psychological Science. APA: American Psychological Association. La idea equivocada 5 —«el ajuste fino o el RLHF son solo un filtro de seguridad desmontable sobre un núcleo neutral»— no aparece explícitamente en estas políticas y, por tanto, no se ilustra aquí. Las letras de las distinciones A–D siguen la Tabla 1.

Políticas de SAGE: https://www.sagepub.com/journals/publication-ethics-policies/artificial-intelligence-policy y https://www.sagepub.com/about/sage-policies/corporate-policies/ai-author-guidelines .

Política de APS: https://www.psychologicalscience.org/publications/aps-editorial-policies .

Políticas de APA: https://www.apa.org/pubs/journals/resources/publishing-policies y https://www.apa.org/pubs/journals/resources/publishing-tips/policy-generative-ai .

Figura 1. Modelo mínimo de trabajo de un sistema basado en LLM



Izquierda: el entrenamiento fuera de línea optimiza un único conjunto de pesos del modelo sobre grandes corpus de texto y multimodales mediante la entropía cruzada del siguiente token, seguido de ajuste por instrucciones y/o RLHF que modifican los mismos parámetros utilizando datos de instrucciones y preferencias. Derecha: durante el despliegue, el modelo de pesos fijos se integra en un producto que gestiona el contexto —memoria de trabajo—, las herramientas/API y la memoria externa; los prompts del usuario, el contenido recuperado y las observaciones de las herramientas se ensamblan en el contexto y se transmiten —opcionalmente a través de filtros de entrada— al LLM, que devuelve una distribución sobre los siguientes tokens. Una política de decodificación —por ejemplo, selección voraz, top-p o muestreo por temperatura—, seguida de filtros de seguridad opcionales, produce una secuencia de salida decodificada que se envía a la interfaz de usuario o se interpreta como una llamada a una herramienta. El diagrama distingue la memoria paramétrica, la contextual y la externa —incluidos los registros que pueden reutilizarse para entrenamiento futuro— y separa la distribución aprendida de las muestras particulares inducidas por el régimen elegido de decodificación y filtrado.

Clave traducida de las etiquetas de la figura original

Entrenamiento (fuera de línea). LLM → Datos de entrenamiento: corpus de texto y multimodales → Preentrenamiento: entropía cruzada del siguiente token; minimizar −log p(siguiente token | contexto) → Ajuste por instrucciones / RLHF: modificación de la política; datos de instrucciones y de preferencias.

Memoria externa: registros; base de conocimiento, perfiles y almacén vectorial. Conexiones: entrenamiento futuro (opcional), consulta de recuperación (opcional) y contenido recuperado.

Herramientas, API y entorno. API: interfaz de programación de aplicaciones. Conexiones: acción o llamada a herramienta (opcional) y observaciones devueltas.

Despliegue (inferencia + sistema). Sistema basado en LLM (capa de integración del producto). Contexto (memoria de trabajo): prompt del usuario + prompt de sistema + historial de conversación → Filtros de entrada (opcionales) → LLM (pesos fijos): memoria paramétrica → Distribución sobre los siguientes tokens: p(· | contexto) → Política de decodificación: selección voraz / top-p / temperatura → Filtros de seguridad → Tokens de salida decodificados: una realización.

Conexiones de salida: añadir tokens al contexto; salida al usuario; acción o llamada a herramienta (opcional). Los trazos discontinuos indican las conexiones o componentes opcionales representados en el original.

Seis ideas equivocadas

Las cuatro distinciones definen un espacio diagnóstico en el que pueden ubicarse seis ideas equivocadas recurrentes: tres sobre estadística y generatividad, dos sobre memoria y alineamiento, y una sobre el estatus cognitivo. Cada una conserva un núcleo de verdad, pero confunde al menos una distinción del modelo mínimo y produce errores característicos en la evaluación, el despliegue o la gobernanza. La Tabla 1 presenta esta lógica como una matriz, que avanza desde la afirmación intuitiva y su núcleo de verdad hacia la distinción confundida, el error resultante, la pregunta correctiva y la orientación práctica. La misma matriz también puede dar cabida a afirmaciones más amplias —por ejemplo, sobre creatividad de tipo humano, inteligencia artificial general, experiencia emocional o apego—, principalmente como variantes de la confusión entre competencia y agencia de la idea equivocada 6.

Ideas equivocadas sobre estadística y generatividad

Idea equivocada 1: «Los LLM son solo predictores del siguiente token» o «loros estocásticos»

En el plano de los mecanismos, los LLM actuales son predictores condicionales del siguiente token entrenados mediante la minimización de la entropía cruzada. Esta descripción, aunque correcta, se vuelve engañosa cuando se la eleva a una explicación completa de lo que son los sistemas basados en LLM o se la utiliza para descartarlos como cognitivamente triviales. Las preocupaciones sociotécnicas que motivan el encuadre de los «loros estocásticos» —el anclaje en el mundo, la procedencia de los datos y la escala— no implican la inferencia adicional de que el objetivo del preentrenamiento delimite la competencia del sistema desplegado (44). El ajuste por instrucciones y el RLHF modifican la distribución aprendida. Una vez que un modelo se integra en un bucle que utiliza herramientas, la predicción del siguiente token puede implementar una política de acciones, cuyas observaciones se incorporan a predicciones posteriores. GeneGPT, que amplía Codex con las API web del NCBI para tareas de genómica, obtuvo 0,83 en la evaluación GeneTuring, frente a 0,00–0,44 de los LLM sin herramientas adicionales (45). El predictor no cambia; el sistema sí. Por ello, las afirmaciones sobre capacidades y riesgos deberían especificar el sistema desplegado —su capa de integración, sus herramientas y su política de decodificación—, y no solo el objetivo de preentrenamiento.

Idea equivocada 2: «Los LLM regresan a la media» o son «el promedio de internet»

Aunque el preentrenamiento por máxima verosimilitud ajusta estructuras de alta probabilidad del corpus de entrenamiento, el modelo aprende una distribución condicional, no una única «respuesta promedio». Esta idea equivocada confunde la distribución aprendida —incluidos los patrones de cola larga y las estructuras raras pero importantes— con las muestras producidas bajo determinados regímenes de decodificación, formulación de prompts y alineamiento, y así trata la homogeneización de las salidas como algo matemáticamente inevitable. El Recuadro 1 destaca lo contrario: los sistemas generativos combinados con búsqueda, aprendizaje por refuerzo, evaluación o selección humana pueden alcanzar regiones de baja probabilidad o poco exploradas y, en algunos dominios, ir más allá de las prácticas humanas históricas en los juegos, el arte y la ciencia. En la práctica, las respuestas insípidas y de tono consensual suelen reflejar decisiones de producto: baja temperatura o decodificación voraz, presiones de alineamiento que favorecen formulaciones genéricas y prompts uniformes de asistente (29, 30). En FreshQA, los prompts aumentados por recuperación mejoraron la exactitud factual de GPT-4 entre un 32,6 % y un 49,0 % en términos relativos respecto del mismo modelo sin recuperación (60); en LitQA2 —una evaluación de literatura científica para expertos de dominio—, PaperQA2, un agente que recupera y sintetiza artículos completos, alcanzó una precisión del 85,2 %, frente al 73,8 % de expertos humanos de nivel doctoral con acceso completo a internet y hasta una semana por conjunto de preguntas (61). En ambos ejemplos, los cambios de rendimiento reflejan la recuperación y la orquestación en torno al mismo modelo paramétrico. La Tabla 1 introduce un diagnóstico sencillo: variar los prompts y los ajustes de decodificación, e informar qué regímenes utilizan efectivamente las evaluaciones y los despliegues.

Idea equivocada 3: «Los LLM solo reproducen los datos de entrenamiento»

Los modelos de lenguaje pueden reproducir fragmentos de su corpus de entrenamiento, en especial textos estandarizados o pasajes célebres, lo que suscita preocupaciones sobre privacidad y derechos de autor (62). Pero tratar a un LLM como una gigantesca tabla de consulta confunde la compresión paramétrica, el uso de prompts contextuales y el almacenamiento externo, y fomenta la suposición de que toda salida es texto copiado. En la práctica, la mayoría de las generaciones no son copias, sino recombinaciones de patrones aprendidos a partir de muchas fuentes. La Tabla 1 reformula la cuestión como una pregunta empírica —¿cuándo las salidas son efectivamente copiadas o casi literales?— y vincula las medidas de mitigación con la selección y el cuidado de los datos, la auditoría de filtraciones y el diseño de la memoria a nivel de producto.

Ideas equivocadas sobre memoria y alineamiento

Idea equivocada 4: «Los LLM aprenden de mis chats y me recuerdan» frente a «Los LLM no tienen estado y no recuerdan nada»

Un polo trata al asistente de chat como un diario en continuo crecimiento que aprende de las conversaciones y recuerda los secretos del usuario; el otro considera que cada respuesta se genera de manera aislada y reduce la continuidad a un truco de la interfaz. Ambos descuidan las tres capas de memoria de los sistemas desplegados. La memoria paramétrica es información codificada en los pesos como una compresión fija y con pérdida de los datos de entrenamiento; aquí, «sin estado» solo significa que esos pesos permanecen fijos durante la inferencia, es decir, durante el uso corriente (28). La memoria contextual es el historial de conversación, el contenido recuperado y otros materiales de entrada que el producto incluye en el prompt actual, y que el modelo solo puede «recordar» mientras permanezcan dentro de la ventana de contexto. La memoria externa, propia del producto —registros, perfiles, documentos e índices de recuperación—, se sitúa fuera del modelo y se reincorpora selectivamente a los prompts; allí se concentran muchos riesgos de privacidad y gobernanza (36). La Tabla 1 desplaza el foco analítico desde «¿el modelo me recuerda?» hacia «¿dónde se almacena, retiene y reutiliza esta información, y bajo qué política?». También sitúa las salvaguardas pertinentes en capas de memoria explícitas y auditables en cada nivel.

Idea equivocada 5: «El ajuste fino y el RLHF son solo filtros superficiales sobre un núcleo neutral»

Una visión frecuente imagina un modelo base neutral en términos de valores, recubierto por una delgada «capa de alineamiento» desmontable, de modo que el ajuste por instrucciones y el RLHF simplemente añaden censura, negativas o cortesía sin modificar el sistema neutral subyacente. En realidad, el ajuste fino supervisado y el RLHF actualizan los mismos parámetros que codifican los conocimientos y las habilidades del modelo, modifican la distribución condicional y producen una política diferente, en lugar de un núcleo intacto con una máscara (30). Tampoco existe una base neutral en términos de valores que pueda recuperarse: el preentrenamiento ya refleja las normas, omisiones y sesgos de muestreo de sus datos, y los procedimientos de alineamiento inscriben además preferencias institucionales en el comportamiento (1). Existen filtros y clasificadores externos que, en ocasiones, pueden añadirse o retirarse sin reentrenamiento; sin embargo, confundirlos con el RLHF fomenta expectativas infundadas sobre la reversibilidad y oculta que el ajuste fino puede introducir compromisos en la cobertura, el estilo y los patrones de rechazo. El modelo de pesos abiertos DeepSeek-R1 ilustra esta arquitectura (63): la censura se implementa en parte en los pesos y en parte mediante filtros durante el despliegue (Recuadro 2). Por eso, la Tabla 1 trata el ajuste fino como un entrenamiento sustantivo y pregunta cómo diverge el comportamiento entre las variantes base y ajustadas cuando los filtros externos se mantienen constantes.

Ideas equivocadas sobre el estatus cognitivo

Idea equivocada 6: «Los LLM piensan como los humanos» frente a «Los LLM no comprenden nada»

El encuadre antropomórfico trata a un LLM como un sujeto unificado con creencias, objetivos y sentimientos estables, una impresión reforzada por interfaces y procedimientos de alineamiento que promueven respuestas en primera persona y socialmente tranquilizadoras. La misma confusión entre competencia y agencia aparece en afirmaciones más amplias según las cuales los LLM actuales ya son inteligencia artificial general, son creativos en el sentido humano o pueden experimentar emociones o apegos. Estas afirmaciones difieren en su contenido, pero todas tratan el desempeño en múltiples tareas, la fluidez estilística, la producción creativa o la interacción relacional como pruebas de agencia, experiencia o comprensión de tipo humano (37, 66). El encuadre deflacionario comete el error simétrico: toma la predicción del siguiente token y la falta de corporalidad como pruebas de que el sistema se limita a manipular símbolos sin acceso al significado. Ambas posiciones pasan por alto que los modelos actuales muestran una comprensión funcional considerable en dominios acotados —generalización sistemática, aprendizaje en contexto y razonamiento de varios pasos—, aunque carecen de corporalidad, aprendizaje persistente en línea y un estado interno estable del tipo que permitiría una comprensión experiencial semejante a la humana (28, 37, 67). Por eso, la Tabla 1 trata esta idea equivocada sobre el estatus cognitivo como una confusión entre competencia y agencia y desplaza la pregunta desde «¿tiene una mente?» hacia «¿qué tareas puede realizar de manera fiable y qué formas de agencia se le atribuyen implícitamente?». Describir la competencia en términos funcionales, en lugar de inferir agencia o estatus moral a partir de un estilo fluido, ayuda a evitar tanto la atribución excesiva como la insuficiente (39).

Implicaciones para la evaluación, el despliegue y la gobernanza

La taxonomía tiene tres objetivos prácticos: la evaluación de capacidades, el diseño del despliegue y la gobernanza institucional.

Evaluación y modelización de capacidades

La evaluación de capacidades debería tomar como unidad de análisis la configuración desplegada, y no el modelo base ni una muestra de chat con ajustes predeterminados. Esto evita dos errores simétricos: las posiciones deflacionarias pasan por alto el comportamiento en bucle cerrado, el uso de herramientas y el aprendizaje en contexto, mientras que las posiciones antropomórficas generalizan en exceso desde demostraciones llamativas hacia una competencia estable. Por ello, las evaluaciones deberían especificar los prompts, los ajustes de decodificación, los sistemas de recuperación, las herramientas, las capas de memoria y los flujos de trabajo bajo los que aparece cada capacidad, poner a prueba su estabilidad ante perturbaciones y examinar los fallos en condiciones de uso realistas.

Un ejemplo de diagnóstico clínico ilustra el argumento. En un ensayo aleatorizado de 2024, los médicos con acceso a GPT-4 obtuvieron solo 2 puntos porcentuales más que un grupo de control con recursos convencionales en razonamiento diagnóstico —una diferencia no significativa—, aunque GPT-4 por sí solo superó al grupo de control en 16 puntos (68). Un ensayo posterior, que utilizó las mismas viñetas y la misma rúbrica de puntuación, pero rediseñó el flujo colaborativo —evaluaciones independientes del profesional y de la IA seguidas de una síntesis generada por IA—, elevó la exactitud de los médicos al 82–85 %, frente al 75 % obtenido con recursos convencionales (69). El modelo no cambió; el flujo de trabajo sí.

Recuadro 1. Distintas vías por las que los sistemas de IA se apartan de las producciones humanas más frecuentes

Las afirmaciones de que los LLM «regresan a la media» o «son el promedio de internet» pasan por alto un punto más amplio: los sistemas de IA que aprenden, especialmente cuando se integran en bucles de búsqueda, evaluación o selección, no tienen por qué limitarse a reproducir las producciones humanas más frecuentes; pueden explorar regiones escasamente representadas en el registro histórico o difíciles de explorar para personas sin asistencia. Los casos siguientes ilustran tres mecanismos —búsqueda guiada por un evaluador, generación de alta variabilidad con filtrado humano y síntesis no limitada por la corporalidad humana—, cada uno de los cuales produce una forma distinta de apartarse de las producciones humanas típicas.

1. Búsqueda guiada por un evaluador en dominios formales

FunSearch combina un LLM preentrenado con un evaluador sistemático para descubrir nuevas construcciones matemáticas (46). AlphaGeometry resuelve problemas de geometría de nivel olímpico mediante la combinación de un modelo de lenguaje neuronal y un motor de deducción simbólica, utilizando datos sintéticos para sortear la escasez de demostraciones humanas (47). Aplicado a 67 problemas de análisis matemático, combinatoria, geometría y teoría de números, AlphaEvolve —un sistema de búsqueda evolutiva guiada por LLM— redescubrió las mejores soluciones conocidas en la mayoría de los casos y mejoró varias otras (48). AlphaGo, aunque no es un modelo de lenguaje, ilustra la misma lógica guiada por un evaluador: el aprendizaje por refuerzo mediante juego contra sí mismo, bajo reglas explícitas, llevó su política hacia regiones del espacio del go escasamente visitadas en el registro humano (49); AlphaGo Zero eliminó por completo los datos humanos (50). La jugada 37 de AlphaGo contra Lee Sedol —legal, muy eficaz y considerada de antemano extraordinariamente improbable para un profesional— se convirtió en un ejemplo canónico: una jugada fuera de la distribución humana previa que luego se incorporó al juego profesional. En estos casos, la novedad no surge del modelo por sí solo, sino de la búsqueda guiada por un evaluador o verificador explícito. Por ello, esta estrategia es más potente cuando las salidas candidatas pueden puntuarse o comprobarse con criterios explícitos y más débil cuando los criterios de la tarea son ambiguos o controvertidos.

2. Generación de alta variabilidad con filtrado humano

Para su exposición basada en DALL·E, Bennett Miller generó más de 100.000 imágenes y seleccionó aproximadamente 20 para exhibirlas (51). Mediante prompts detallados, revisiones iterativas y una selección rigurosa, trató al modelo como un generador de alta variabilidad y al juicio humano como un filtro. El trabajo empírico a gran escala respalda esta explicación centrada en el flujo de trabajo. En un conjunto de más de 4 millones de obras de arte, los creadores asistidos por IA que combinaron una ideación activa con el filtrado selectivo de las salidas del modelo recibieron las evaluaciones más favorables de sus pares; aunque disminuyeron la novedad media del contenido y la novedad visual, aumentó la novedad máxima del contenido entre los creadores que exploraron con éxito el espacio de ideas (52). Un estudio posterior de 31.076 creadores también encontró que la frontera de ideas se expandió en términos absolutos gracias al aumento de la producción, aunque descendieron las tasas de creatividad histórica —H-creativity— por pieza y no se detectó un efecto de complementariedad humano-IA más allá de las ganancias de productividad (53). Sin embargo, el uso predeterminado o restringido a un dominio puede homogeneizar la producción: los relatos asistidos por LLM se juzgan más creativos individualmente, pero se vuelven más similares entre sí (54); las imágenes de DALL·E pueden exhibir una «singularidad genérica», con una diversidad superficial sostenida por patrones estandarizados de representación (55); y un estudio de corpus encontró que las imágenes de la guerra entre Rusia y Ucrania generadas por IA presentaban una versión depurada del conflicto al excluir la muerte, las lesiones y el sufrimiento de niños y refugiados, mientras sobrerrepresentaban las escenas urbanas (56). En conjunto, estos estudios sitúan tanto la divergencia respecto de la producción promedio como la convergencia hacia ella en el nivel del flujo de trabajo sociotécnico, y no solo en el modelo: el modelo aporta variabilidad; los ajustes predeterminados y las restricciones compartidas la reducen; el volumen amplía el conjunto de candidatos; y el filtrado humano determina el resultado.

3. Síntesis no limitada por la corporalidad humana

Los sistemas de generación de audio y video pueden sintetizar interpretaciones que no están limitadas por la fisiología humana, las condiciones de producción en vivo o la frecuencia histórica de los estilos registrados. Los sistemas de generación musical, por ejemplo, pueden producir frases vocales extensas sin pausas para respirar e híbridos de géneros o timbres raros o ausentes en los corpus grabados (57–59). La generación de video y la edición asistida por IA extienden la misma lógica a la acción: gestos exagerados, movimientos de cámara imposibles, transformaciones fantásticas, secuencias de acción estilizadas y maniobras biomecánicamente difíciles o inverosímiles pueden generarse sin la misma dependencia de intérpretes en vivo, decorados, equipos de especialistas, animación o cadenas convencionales de efectos visuales. Esto desplaza parte del flujo de trabajo desde la filmación y la interpretación hacia la generación, la selección y la edición.

Despliegue y diseño de sistemas

Las confusiones sobre la memoria producen arquitecturas frágiles y, a veces, inseguras: los sistemas pueden suponer una continuidad de la que el modelo carece o almacenar y reutilizar datos de usuarios en memorias opacas a nivel de producto. Distinguir la memoria paramétrica, la contextual y la externa aclara qué está fijado en los pesos, qué puede permanecer en un contexto de corta duración y qué corresponde a un almacenamiento explícito y auditable con controles de acceso separados. Del mismo modo, abandonar el mito del «núcleo neutral más filtro» obliga a los diseñadores a tratar el ajuste fino y el RLHF como intervenciones sustantivas y cargadas de valores cuyos compromisos en cobertura, estilo y patrones de rechazo deben medirse y, cuando sea posible, volverse reversibles o, al menos, auditables. El caso de DeepSeek-R1 (Recuadro 2) muestra cómo un mismo punto de partida de pesos abiertos puede integrarse, alinearse, filtrarse o realinearse en regímenes normativos divergentes, según quién controle el posentrenamiento y el despliegue.

Gobernanza y razonamiento público

En los debates sobre gobernanza, tratar a los LLM como simples loros subestima los riesgos derivados de su falta de fiabilidad, su escala y su uso indebido; tratarlos como protopersonas desvía el debate hacia preguntas prematuras sobre estatus moral o «derechos de la IA». El modelo mínimo, en cambio, dirige el escrutinio hacia mecanismos específicos de diseño: la selección y el cuidado de los datos de entrenamiento, los objetivos de alineamiento y la supervisión de los datos de preferencias, las interfaces de herramientas y memoria, y los ajustes predeterminados de decodificación y prompts.

Estas distinciones ya configuran decisiones jurídicas y políticas relevantes, aunque de manera desigual. En Moffatt contra Air Canada, un tribunal responsabilizó a la aerolínea por la información errónea de su chatbot y rechazó la sugerencia de que este pudiera tener una responsabilidad separada; la rendición de cuentas correspondió a los operadores del sistema desplegado. Después de que Mata contra Avianca expusiera citas de casos judiciales inventadas por IA, los tribunales emitieron órdenes de declaración y certificación, pero algunas regulan el «uso de IA» en abstracto en lugar de abordar el riesgo específico de fuentes fabricadas (70, 71). El Perfil de IA Generativa del NIST ofrece un enfoque más sensible a la arquitectura (72): distingue sistemas preentrenados, adaptados y desplegados, y solicita documentar el ajuste fino, el aumento por recuperación y el seguimiento posterior al despliegue: las mismas distinciones que hace explícitas la Figura 1.

Las políticas editoriales sobre IA como casos de estudio de gobernanza

La publicación científica constituye un caso útil de gobernanza porque las concepciones abstractas sobre los LLM se traducen en reglas para autores, revisores y editores. Estas políticas están muy extendidas, pero presentan grados desiguales de especificación y, tal como están redactadas actualmente, sus efectos observados son limitados. Las principales editoriales y revistas difieren en lo que consideran «uso de IA» y en cómo debe declararse ese uso (73, 74). En un análisis a nivel de revistas, aproximadamente el 70 % había adoptado políticas sobre IA, en su mayoría requisitos de declaración; sin embargo, la escritura asistida por IA siguió aumentando, las declaraciones explícitas continuaron siendo poco frecuentes y no se detectaron diferencias en las tasas de escritura asistida por IA entre revistas con políticas y revistas sin ellas (75).

La Tabla 2 aplica la matriz diagnóstica de la Tabla 1 a una selección de políticas actuales sobre IA de grandes editoriales y asociaciones de psicología. Estas políticas enfatizan, acertadamente, la declaración del uso, la verificación de citas y la confidencialidad, pero el lenguaje con el que las justifican también desdibuja las distinciones de la Figura 1. Surgen tres patrones. Primero, los pasajes sobre fiabilidad epistémica suelen adoptar encuadres estadísticos deflacionarios: vinculan directamente el texto generado por IA o las fechas de corte del entrenamiento con la falta de fiabilidad científica, como si el objetivo de predecir el siguiente token o la fecha de corte determinaran por sí solos la calidad epistémica. Este razonamiento reproduce los errores de «solo estadística» de las ideas equivocadas 1 y 2 al ignorar la recuperación, el uso de herramientas y la verificación. Las advertencias relacionadas con la reproducción de grandes segmentos de texto reproducen la visión de la memorización como tabla de consulta de la idea equivocada 3, mientras prestan menos atención a la apropiación mediante paráfrasis o de estructuras.

Segundo, los pasajes sobre memoria y confidencialidad suelen tratar los materiales introducidos en sistemas de IA generativa como inherentemente accesibles a los proveedores, confundiendo así la memoria paramétrica, la contextual y la externa —idea equivocada 4— y los diferentes regímenes de despliegue. En lugar de centrarse en los flujos y las políticas de datos, tratan a la IA generativa como un régimen unitario de manejo de datos, y fusionan en un único riesgo indiferenciado despliegues con propiedades sustancialmente distintas de almacenamiento, registro y reutilización.

Tercero, las afirmaciones de que los LLM no pueden reproducir el «pensamiento creativo y crítico humano» incorporan la idea equivocada 6 al lenguaje de las políticas al introducir una tesis cognitiva fuerte en el discurso de gobernanza, en lugar de vincular las restricciones con la competencia a nivel de tareas, la fiabilidad documentada y el uso responsable. En cambio, la idea equivocada 5 está prácticamente ausente de estas políticas: dicen poco sobre el ajuste fino, el RLHF o cómo interactúa el alineamiento inscrito en los pesos con los filtros externos.

Recuadro 2. La censura en DeepSeek-R1 como caso de estudio de alineamiento y despliegue

DeepSeek-R1 es un modelo de pesos abiertos cuyo comportamiento depende fuertemente del alineamiento posterior al entrenamiento y de las decisiones de despliegue (63–65).

Entrenamiento, alineamiento y filtros

R1 primero fue preentrenado, luego ajustado por instrucciones y alineado mediante RLHF bajo restricciones impuestas por el Estado. En la aplicación oficial, el modelo alineado está rodeado por filtros del servidor que supervisan los prompts y sus continuaciones, interrumpiendo o sobrescribiendo respuestas sobre temas políticamente sensibles, como las protestas de Tiananmén de 1989 o el estatus político de Taiwán. Los pesos subyacentes permiten un sólido desempeño en matemáticas y programación, pero las consultas políticas en ese despliegue producen negativas o respuestas ajustadas a la línea del partido, y los prompts mixtos pueden hacer que el modelo abandone el razonamiento paso a paso en cuanto aparece una subpregunta sensible.

Pesos abiertos y «realineamiento»

Dado que los pesos están disponibles públicamente, grupos independientes han modificado la política de comportamiento de R1 sin reentrenarlo desde cero. Una línea de trabajo ajusta R1 con respuestas factuales y sin censura a preguntas antes suprimidas, produciendo variantes que conservan la capacidad de razonamiento y reducen las negativas políticas. Otra utiliza la edición selectiva de pesos o la compresión para identificar y eliminar los parámetros más asociados con la censura, e informa de modelos que responden preguntas políticamente sensibles sin negarse, al tiempo que conservan en gran medida su rendimiento en las evaluaciones.

Lecciones

La censura en R1 no es una delgada máscara desmontable sobre un núcleo neutral ni una esencia inmutable del modelo. Es un componente sustancial, pero parcialmente editable, de la política aprendida, complementado por filtros durante el despliegue. El caso cristaliza tres distinciones: entre el preentrenamiento y el alineamiento posterior; entre el comportamiento configurado en los pesos y el impuesto por filtros externos; y entre un sistema paramétrico y los múltiples regímenes normativos que pueden instanciarse a partir de él. Los pesos abiertos de DeepSeek-R1 hacen que esta descomposición sea más abordable; en los sistemas de pesos cerrados, la frontera entre el comportamiento a nivel de pesos y el filtrado externo es empíricamente opaca.

Conclusión

En conjunto, el modelo mínimo de trabajo y la taxonomía de seis ideas equivocadas ofrecen un conjunto de herramientas diagnósticas para las afirmaciones sobre los LLM. Al evaluar lo que se afirma sobre qué son estos sistemas, qué pueden hacer o qué implican, el marco plantea tres preguntas: qué nivel de descripción se invoca; qué distinción se está confundiendo; y qué consecuencias se desprenden para la evaluación, el despliegue y la gobernanza si esa distinción se mantiene. Aunque no resuelve si los LLM realmente comprenden, razonan, poseen agencia o merecen estatus moral, reduce el riesgo de que las políticas y las prácticas sean orientadas por consignas, en lugar de por la arquitectura, el comportamiento y las condiciones de despliegue de los sistemas en cuestión.

Agradecimientos

Se utilizó GPT-5.1 Pro para revisar el manuscrito, siguiendo los prompts descritos en https://www.nature.com/articles/s41551-024-01185-8 .

Conflictos de intereses

El autor declara no tener conflictos de intereses.

Financiación

El autor recibió apoyo del proyecto Brain Science and Brain-like Intelligence Technology—National Science and Technology Major Project (2021ZD0204200), del Fondo de Investigación de la Universidad Yonsei (5355661) y de la beca Yonsei Fellowship, financiada por Lee Youn Jae.

Contribuciones del autor

Zhicheng Lin: conceptualización; visualización; redacción del borrador original; revisión y edición del texto.

Disponibilidad de datos

No hay datos subyacentes a este trabajo.

Referencias

Se conservan la numeración y los datos bibliográficos del artículo original. Los títulos permanecen en su idioma de publicación.

1. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the dangers of stochastic parrots: can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency; Virtual Event, Canada. Association for Computing Machinery; 2021. p. 610–623.

2. Hicks MT, Humphries J, Slater J. 2024. ChatGPT is bullshit. Ethics Inf Technol. 26:38.

3. Chiang T. 2023. ChatGPT is a blurry JPEG of the web. In The New Yorker. https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web (accessed 21 December 2025).

4. Brodeur PG, et al. 2024. Superhuman performance of a large language model on the reasoning tasks of a physician. arXiv 10849. https://doi.org/10.48550/arXiv.2412.10849, preprint: not peer reviewed.

5. Wang L, et al. 2024. A survey on large language model based autonomous agents. Front Comput Sci (Berl). 18:186345.

6. Bubeck S, et al. 2023. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv 12712. https://doi.org/10.48550/arXiv.2303.12712, preprint: not peer reviewed.

7. Schoenegger P, Tuminauskaite I, Park PS, Bastos RVS, Tetlock PE. 2024. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. Sci Adv. 10:eadp1528.

8. DeVito MA, Gergle D, Birnholtz J. "Algorithms ruin everything": #RIPTwitter, folk theories, and resistance to algorithmic change in social media. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems; Denver, Colorado, USA. Association for Computing Machinery; 2017. p. 3163–3174.

9. Gelman SA, Legare CH. 2011. Concepts and folk theories. Annu Rev Anthropol. 40:379–398.

10. Bewersdorff A, Zhai X, Roberts J, Nerdel C. 2023. Myths, mis-and preconceptions of artificial intelligence: a review of the literature. Comput Educ Artif Intell. 4:100143.

11. Cheng M, Gligorić K, Piccardi T, Jurafsky D. AnthroScore: a computational linguistic measure of anthropomorphism. Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, St. Julian's, Malta, 2024. p. 807–825.

12. Ryazanov I, Öhman C, Björklund J. 2025. How ChatGPT changed the media's narratives on AI: a semi-automated narrative analysis through frame semantics. Minds Mach (Dordr). 35:2.

13. Lee H, Park H. 2026. Framing generative AI: an analysis of impact, quoted sources, and attributions of responsibilities in U.S. elite media. J Mass Commun Q. https://doi.org/10.1177/10776990251410593

14. Cheng M, et al. 2026. Metaphors of AI indicate that people increasingly perceive AI as warm and human-like. Commun Psychol. 4:8.

15. Brachman M, et al. Building appropriate mental models: what users know and want to know about an agentic AI chatbot. Proceedings of the 30th International Conference on Intelligent User Interfaces; Cagliari, Italy. Association for Computing Machinery, 2025. p. 247–264.

16. Jones B, Stemmler K, Su E, Kim Y-H, Kuzminykh A. Users' expectations and practices with agent memory. Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems; Yokohama, Japan. Association for Computing Machinery, 2025. p. Article 563.

17. Cohn M, et al. Believing anthropomorphism: examining the role of anthropomorphic cues on trust in large language models. Extended Abstracts of the CHI Conference on Human Factors in Computing Systems; Honolulu, HI, USA. Association for Computing Machinery; 2024. p. Article 54.

18. Inie N, Druga S, Zukerman P, Bender EM. From “AI” to probabilistic automation: how does anthropomorphization of technical systems descriptions influence trust? Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency; Rio de Janeiro, Brazil. Association for Computing Machinery; 2024. p. 2322–2347.

19. Li Z, Kim N, Lou C. 2026. Into the black box: laypeople’s folk theories about generative artificial intelligence chatbots. Big Data Soc. 13:20539517261447838.

20. Heinsfeld BD, Veletsianos G. 2025. The language on GenAI: a critical exploration of personification metaphors in UNESCO’s guidance for generative AI in education and research. J Interact Media Educ. 2025:16.

21. Bommasani R, et al. 2021. On the opportunities and risks of foundation models. arXiv 07258. https://doi.org/10.48550/arXiv.2108.07258, preprint: not peer reviewed.

22. Weidinger L, et al. Taxonomy of risks posed by language models. Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency; Seoul, Republic of Korea. Association for Computing Machinery; 2022. p. 214–229.

23. Shanahan M. 2024. Talking about large language models. Commun ACM. 67:68–79.

24. Peter S, Riemer K, West JD. 2025. The benefits and dangers of anthropomorphic conversational agents. Proc Natl Acad Sci U S A. 122:e2415898122.

25. Zhang D, et al. 2025. Memory in large language models: mechanisms, evaluation and evolution. arXiv 18868. https://doi.org/10.48550/arXiv.2509.18868, preprint: not peer reviewed.

26. Zhang Z, et al. 2025. A survey on the memory mechanism of large language model-based agents. ACM Trans Inf Syst. 43: Article 155.

27. Vaswani A, et al. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA. Curran Associates Inc.; 2017. p. 6000–6010.

28. Brown T, et al. 2020. Language models are few-shot learners. Adv Neural Inf Process Syst. 33:1877–1901.

29. Holtzman A, Buys J, Du L, Forbes M, Choi Y. 2020. The curious case of neural text degeneration. International Conference on Learning Representations (ICLR 2020).

30. Ouyang L, et al. 2022. Training language models to follow instructions with human feedback. Adv Neural Inf Process Syst. 35:27730–27744.

31. Christiano PF, et al. 2017. Deep reinforcement learning from human preferences. Adv Neural Inf Process Syst. 30: 4299–4307.

32. Ziegler DM, et al. 2019. Fine-tuning language models from human preferences. arXiv 08593. https://doi.org/10.48550/arXiv. 1909.08593, preprint: not peer reviewed.

33. Schick T, et al. 2023. Toolformer: language models can teach themselves to use tools. Adv Neural Inf Process Syst. 36: 68539–68551.

34. Yao S, et al. 2023. ReAct: synergizing reasoning and acting in language models. The Eleventh International Conference on Learning Representations (ICLR 2023); Kigali, Rwanda. OpenReview.net.

35. Nakano R, et al. 2021. WebGPT: browser-assisted question-answering with human feedback. arXiv 09332. https://doi.org/10.48550/arXiv.2112.09332, preprint: not peer reviewed.

36. Lewis P, et al. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv Neural Inf Process Syst. 33:9459–9474.

37. Lin Z. 2025. Six fallacies in substituting large language models for human participants. Adv Meth Pract Psychol Sci. 8: 25152459251357566.

38. Wu Y, et al. 2025. From human memory to AI memory: a survey on memory mechanisms in the era of LLMs. arXiv 15965. https://doi.org/10.48550/arXiv.2504.15965, preprint: not peer reviewed.

39. Lin Z. 2026. Large language models as psychological simulators: a methodological guide. Adv Meth Pract Psychol Sci. 9: 25152459251410153.

40. Lin Z. 2026. A validity-guided workflow for robust large language model research in psychology. Behav Res Methods. 58:216.

41. Lin Z. From prompts to constructs: a dual-validity framework for large language model research in psychology. Annu Rev Psychol. https://doi.org/10.1146/annurev-psych-100925-0348 07

42. Norman DA. Some observations on mental models. In: Gentner D, Stevens AL, editors. Mental models. Lawrence Erlbaum Associates, 1983. p. 7–14.

43. Weisberg M. 2007. Three kinds of idealization. J Philos. 104: 639–659.

44. Bender EM, Costello E, Lee K, Farrow R, Ferreira G. 2025. Unsafe AI for education: a conversation on stochastic parrots and other learning metaphors. J Interact Media Educ. 2025: 10.

45. Jin Q, Yang Y, Chen Q, Lu Z. 2024. GeneGPT: augmenting large language models with domain tools for improved access to biomedical information. Bioinformatics. 40:btae075.

46. Romera-Paredes B, et al. 2024. Mathematical discoveries from program search with large language models. Nature. 625:468–475.

47. Trinh TH, Wu Y, Le QV, He H, Luong T. 2024. Solving Olympiad geometry without human demonstrations. Nature. 625: 476–482.

48. Georgiev B, Gómez-Serrano J, Tao T, Wagner AZ. 2025. Mathematical exploration and discovery at scale. arXiv 02864. https://doi.org/10.48550/arXiv.2511.02864, preprint: not peer reviewed.

49. Silver D, et al. 2016. Mastering the game of Go with deep neural networks and tree search. Nature. 529:484–489.

50. Silver D, et al. 2017. Mastering the game of Go without human knowledge. Nature. 550:354–359.

51. Chiang T. 2024. Why AI isn't going to make art. The New Yorker. https://www.newyorker.com/culture/the-weekend-essay/why-ai-isnt-going-to-make-art (accessed 21 December 2025).

52. Zhou E, Lee D. 2024. Generative artificial intelligence, human creativity, and art. PNAS Nexus. 3:pgae052.

53. Zhou EB, Lee D, Gu B. 2025. Who expands the human creative frontier with generative AI: hive minds or masterminds? Sci Adv. 11:eadu5800.

54. Doshi AR, Hauser OP. 2024. Generative AI enhances individual creativity but reduces the collective diversity of novel content. Sci Adv. 10:eadn5290.

55. Westberg G, Kvåle G. 2025. The generic uniqueness of AI imagery: a critical approach to Dall-E as semiotic technology. Discourse Soc. 36:575–598.

56. Laba N, Roman N, Parmelee JH. 2025. Memory of the multitude and representation in AI-generated images of war. Mem Mind Media. 4:e14.

57. Borsos Z, et al. 2023. AudioLM: a language modeling approach to audio generation. IEEE/ACM Trans Audio Speech Lang Process. 31:2523–2533.

58. Agostinelli A, et al. 2023. MusicLM: generating music from text. arXiv 11325. https://doi.org/10.48550/arXiv.2301.11325, preprint: not peer reviewed.

59. Casini L, Vila LC, Dalmazzo D, Kaila AK, Sturm BLT. 2026. Data-driven analysis of text-conditioned AI-generated music: a case study with Suno and Udio. Trans Int Soc Music Inf Retr. 9:194–209.

60. Vu T, et al. FreshLLMs: refreshing large language models with search engine augmentation. Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, Bangkok, Thailand, 2024. p. 13697–13720.

61. Skarlinski MD, et al. 2024. Language agents achieve superhuman synthesis of scientific knowledge. arXiv 13740. https://doi.org/10.48550/arXiv.2409.13740, preprint: not peer reviewed.

62. Carlini N, et al. Extracting training data from large language models. 30th USENIX Security Symposium (USENIX Security 21); Virtual Event. USENIX Association, 2021. p. 2633–2650.

63. Guo D, et al. 2025. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. Nature. 645:633–638.

64. Naseh A, et al. 2025. R1dacted: investigating local censorship in DeepSeek's R1 language model. arXiv 12625. https://doi. org/10.48550/arXiv.2505.12625, preprint: not peer reviewed.

65. Yang Z. 2025. Here's how DeepSeek censorship actually works—and how to get around it. In Wired. https://www. wired.com/story/deepseek-censorship (accessed 21 December 2025).

66. Colombatto C, Birch J, Fleming SM. 2025. The influence of mental state attributions on trust in large language models. Commun Psychol. 3:84.

67. Wei J, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. Adv Neural Inf Process Syst. 35: 24824–24837.

68. Goh E, et al. 2024. Large language model influence on diagnostic reasoning: a randomized clinical trial. JAMA Netw Open. 7:e2440969.

69. Everett SS, et al. 2026. From tool to teammate in a randomized controlled trial of clinician-AI collaborative workflows for diagnosis. NPJ Digit Med. 9:409.

70. Grossman MR, Grimm PW, Brown DG. 2023. Is disclosure and certification of the use of generative AI really necessary? Judicature. 107:69–77.

71. Gunder JR. 2024. Rule 11 is no match for generative AI. Stanf Technol Law Rev. 27:308–361.

72. National Institute of Standards and Technology. Artificial intelligence risk management framework: generative artificial intelligence profile. U.S. Department of Commerce, 2024.

73. Ganjavi C, et al. 2024. Publishers' and journals' instructions to authors on use of generative artificial intelligence in academic and scientific publishing: bibliometric analysis. BMJ. 384: e077192.

74. Lin Z. 2024. Towards an AI policy framework in scholarly publishing. Trends Cogn Sci. 28:85–88.

75. He Y, Bu Y. 2026. Academic journals' AI policies fail to curb the surge in AI-assisted academic writing. Proc Natl Acad Sci U S A. 123:e2526734123.