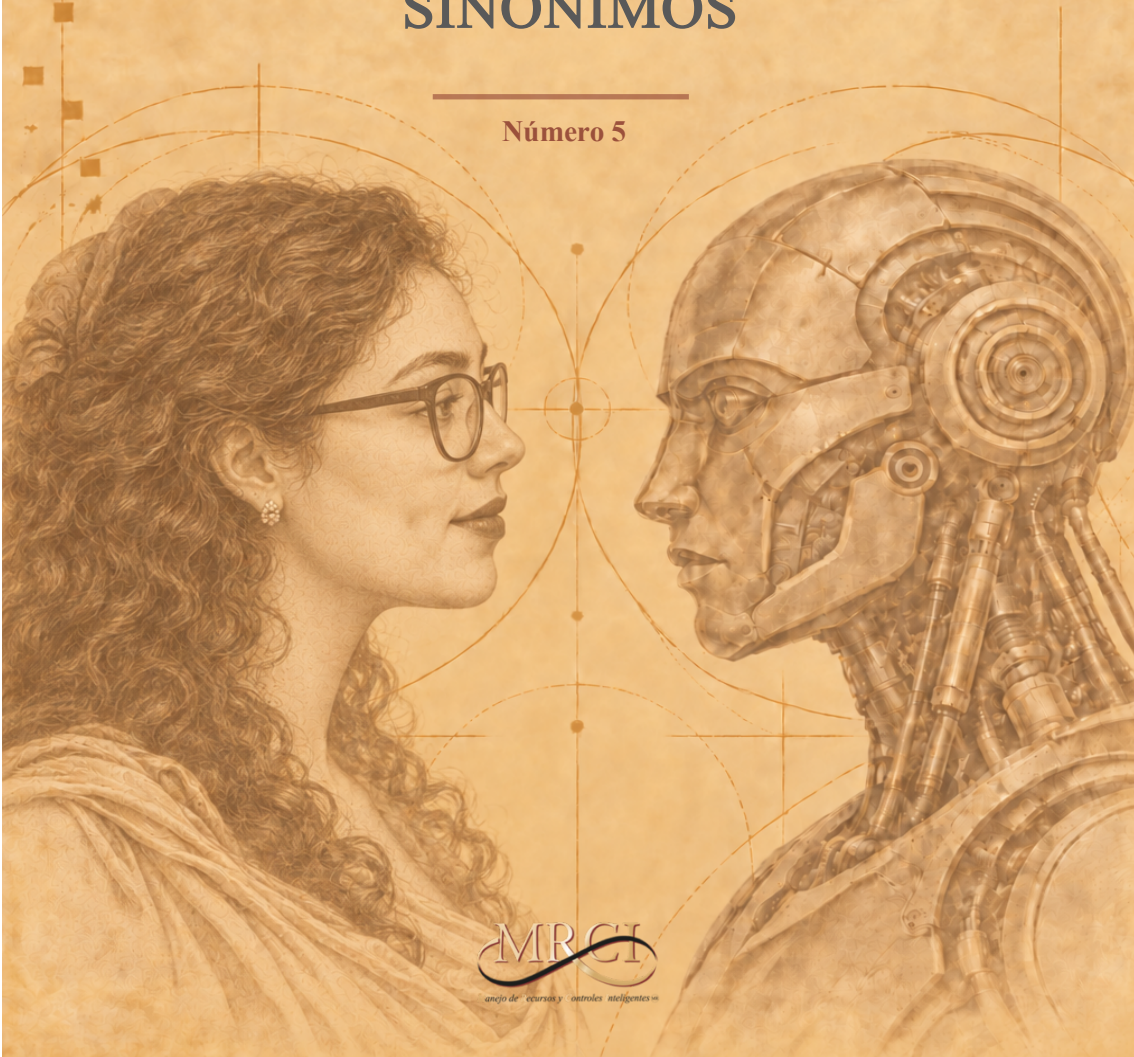
COLECCIÓN

MITOS Y REALIDADES DE LA IA

Entre la promesa y la realidad

IA GENERATIVA Y MODELOS DE LENGUAJE Y POR QUÉ NO SON SINÓNIMOS

Número 5



MRCI

anejo de ecursos y ontroles nteligentes

DR. JOSÉ ANTONIO PÉREZ RAMOS

IA GENERATIVA Y MODELOS DE LENGUAJE Y POR QUÉ NO SON SINÓNIMOS

SOBRE EL AUTOR

Doctor en Ciencias de lo Fiscal por el Instituto de Especialización para Ejecutivos. Maestro en Fiscal por la Universidad Autónoma "Benito Juárez" de Oaxaca. Maestro en Estudios Avanzados en Derechos Humanos por la Global Open University. Licenciado en Contaduría Pública por la Universidad Autónoma "Benito Juárez" de Oaxaca y Licenciado en Filosofía por la Universidad Nacionalista de México.

Socio fundador y director general de la Firma Manejo de Recursos y Controles Inteligentes (Firma MRCI). Fiscalista del año 2009 por la revista *Defensa Fiscal*. Doctor honoris causa por la Asamblea General del Claustro Doctoral Iberoamericano (2016). Doctor honoris causa por el Instituto Mexicano de Líderes de Excelencia (2017). Doctor honoris causa por 1Million Startups, Latinomics, Leaderships Forum y la Fundación Humanist World (2018). Doctor honoris causa por el Claustro de Doctores de la AISOL (American International School of Law) (2019).

Autor de diversas obras, entre las que destacan *Ius mutantur o del derecho que cambia* (Tirant Lo Blanch, 2025), *Futurología* (MRCI, 2025) y *Hasta que la muerte nos separe* (MRCI, 2025). Ha participado como coordinador en *Cuestiones tributarias. Problemas y controversias en el México actual* (Tirant Lo Blanch, 2023), *El costo de la justicia* (Apex Iuris, 2019), *Remuneraciones estratégicas inteligentes* (MRCI, 2015) y *Retos de la justicia tributaria* (Tirant Lo Blanch-AMDF, 2025).

Titular de *Pie de Página*, programa para la divulgación de los derechos humanos en México, a través de la plataforma Legal Cloud. Expositor en foros nacionales e internacionales. Colaborador en diversas revistas jurídicas, entre las que destacan *Defensa Fiscal* y *Legal Cloud*.



IA GENERATIVA Y MODELOS DE LENGUAJE Y POR QUÉ NO SON SINÓNIMOS

DR. JOSÉ ANTONIO PÉREZ RAMOS

PRIMERA EDICIÓN, ATEMPORAL

Derechos reservados, propiedad de
Dr. José Antonio Pérez Ramos

Comentarios y opiniones: investigacion@mrci.com.mx

Título original: IA Generativa y modelos de lenguaje y por qué no son sinónimos.

Colección mitos y realidades de la inteligencia artificial.

Entre la promesa y la realidad.

Autor: Dr. José Antonio Pérez Ramos.

Queda prohibida la reproducción total y parcial de esta obra denominada; IA GENERATIVA Y MODELOS DE LENGUAJE Y POR QUÉ NO SON SINÓNIMOS, por cualquier medio, sin autorización escrita del autor.

PRINTED IN MEXICO
IMPRESO EN MÉXICO

ÍNDICE

INTRODUCCIÓN ..... - 7 -
LA INTERFAZ Y EL MECANISMO ..... - 8 -
MÉTODO Y ALCANCE DE LA REVISIÓN ..... - 9 -
CUATRO EJES DE CLASIFICACIÓN ..... - 10 -
MODELOS DE LENGUAJE ..... - 13 -
UN EJEMPLO MÍNIMO DE GENERACIÓN SECUENCIAL ..... - 15 -
LENGUAJE ENMASCARADO Y REPRESENTACIÓN ..... - 16 -
GENERACIÓN FUERA DEL LENGUAJE ..... - 18 -
REDES GENERATIVAS ADVERSARIAS ..... - 20 -
AUTOENCODIFICADORES VARIACIONALES ..... - 21 -
DIFUSIÓN Y ELIMINACIÓN DE RUIDO ..... - 23 -
SISTEMAS HÍBRIDOS Y MULTIMODALES ..... - 24 -
MULTIMODAL NO SIGNIFICA NECESARIAMENTE GENERATIVO ..... - 27 -
DIFERENCIAS ENTRE FAMILIAS ..... - 28 -
CALIDAD Y COMPROBACIÓN ..... - 31 -
MÉTRICAS ÚTILES CON ALCANCES LIMITADOS ..... - 32 -
UN PROMEDIO PUEDE OCULTAR EL ERROR QUE IMPORTA ..... - 34 -
CASO DE REFERENCIAS GENERADAS ..... - 35 -
CASO DE RECONSTRUCCIÓN VISUAL ..... - 37 -
CASO DE BÚSQUEDA MULTIMODAL ..... - 39 -
NOVEDAD Y MEMORIZACIÓN PUEDEN COEXISTIR ..... - 42 -
EVALUACIÓN SEGÚN LA TAREA ..... - 43 -
OBJECIONES Y LÍMITES DE LA DISTINCIÓN ..... - 47 -
CONCLUSIÓN ..... - 50 -
REFERENCIAS ..... - 51 -

Introducción

El texto explica por qué la inteligencia artificial generativa y los modelos de lenguaje no son sinónimos. Mientras la primera se relaciona con la producción de contenido sintético en modalidades como texto, imagen y audio, los segundos se ocupan del lenguaje como objeto de representación y predicción. Mediante una revisión documental y crítica, distingue cuatro dimensiones para comprender estos sistemas: modalidad, objetivo de aprendizaje, mecanismo de funcionamiento y tarea concreta. Examina modelos lingüísticos, redes generativas adversarias, autoencodificadores variacionales y modelos de difusión, además de sistemas híbridos y multimodales, para mostrar que utilizar instrucciones escritas o combinar texto e imágenes no implica realizar la misma operación.

El documento sostiene que la evaluación debe corresponder al propósito de cada tarea y diferenciar plausibilidad, fidelidad, exactitud y diversidad. A través de casos hipotéticos sobre referencias bibliográficas, reconstrucciones visuales y búsqueda de imágenes, muestra que una respuesta convincente puede contener información falsa, detalles sin respaldo u omisiones relevantes.

La interfaz y el mecanismo

Una persona escribe una instrucción y recibe un ensayo; otra escribe una instrucción semejante y obtiene una imagen, mientras una tercera formula una consulta para localizar documentos ya existentes; las tres experiencias pueden compartir una caja de texto, pero esa semejanza no permite concluir que ejecutan la misma operación ni que sus resultados merecen la misma confianza.

La interfaz oculta diferencias relevantes: redactar una respuesta, ordenar resultados y sintetizar una imagen requieren procesos que pueden colaborar dentro de una aplicación sin volverse equivalentes, por lo que identificar cualquier herramienta con lenguaje natural como un modelo de lenguaje confunde el medio de interacción con el mecanismo que realiza el trabajo.

La pregunta central es qué relación existe entre IA generativa y modelos de lenguaje, y qué consecuencias prácticas tiene describirla con precisión; la tesis de este artículo es que ambas expresiones clasifican propiedades diferentes, de modo que su intersección es importante, pero ninguna autoriza una explicación universal de todos los sistemas incluidos en la otra.

Un error de vocabulario puede convertirse en un error de evaluación: si se cree que generar una respuesta equivale a recuperar información, bastará con que la respuesta parezca pertinente para aceptarla; si se reconoce que se produjo contenido, habrá que comprobar qué partes proceden de una fuente, cuáles son transformaciones y cuáles requieren una justificación independiente.

Método y alcance de la revisión

La investigación es documental y crítica, con corte de consulta al 2 de octubre de 2026; se seleccionaron publicaciones originales por su capacidad para demostrar diferencias conceptuales, explicar mecanismos y examinar límites de evaluación, sin pretender una revisión sistemática de toda la literatura ni una clasificación de productos comerciales disponibles en esa fecha.

Se consultaron trabajos fundacionales junto con investigaciones posteriores que permiten corregir generalizaciones sobre calidad, diversidad y memorización; las fechas de los modelos se conservan como referencias históricas y no se presentan sus resultados como clasificaciones actuales, mientras los



casos y cálculos del artículo son ejemplos hipotéticos de elaboración propia, no experimentos ejecutados.

El alcance excluye una exposición completa de tokenización, atención y escalamiento, asuntos que requieren tratamiento propio; aquí se utiliza únicamente el detalle técnico necesario para responder por qué dos sistemas pueden generar contenido por mecanismos diferentes y por qué una capacidad lingüística o multimodal no implica, por sí sola, generación abierta ni verificación factual.

Cuatro ejes de clasificación

El primer eje es la modalidad u objeto del trabajo: texto, imagen, audio u otras representaciones; el segundo es el objetivo de aprendizaje, es decir, aquello que el entrenamiento intenta mejorar mediante una función de pérdida; el tercero es el mecanismo usado al producir una salida, y el cuarto es la tarea que el sistema realiza en una situación concreta.

Estas dimensiones no tienen una correspondencia de uno a uno: una arquitectura puede admitir distintos objetivos,

un objetivo puede emplearse con varias modalidades y un modelo entrenado para generar puede utilizarse en clasificación; por ello, conviene describir una solución mediante varias propiedades antes de asignarle una etiqueta general.

En el uso contemporáneo de la expresión IA generativa suele destacarse la producción de contenido sintético, pero en estadística la palabra generativo tiene también un significado relativo a la distribución modelada; Ng y Jordan (2001) comparan clasificadores que modelan conjuntamente datos y clases con otros que modelan directamente la clase condicionada a los datos, distinción que no equivale a comparar un chatbot con un clasificador.

Esta precisión histórica evita afirmar que todo procedimiento llamado generativo tiene como producto habitual un texto o una imagen para el público; un clasificador puede apoyarse en supuestos generativos y entregar solamente una etiqueta, de la misma manera que una aplicación puede utilizar un modelo capaz de redactar y limitar su salida a una categoría predeterminada.

La Tabla 1 propone una ficha descriptiva por ejes, cuya finalidad es aclarar preguntas antes de comparar sistemas; no constituye una taxonomía definitiva de todas las arquitecturas, sino una herramienta para detectar cuándo dos afirmaciones hablan de niveles diferentes.

Tabla 1. *Preguntas que describen propiedades distintas de un sistema*

Dimensión	**Pregunta pertinente**	**Ejemplo de respuesta**
Modalidad	¿Qué representaciones recibe y produce?	Recibe texto y produce una imagen
Objetivo	¿Qué relación intenta aprender?	Distinguir parejas de texto e imagen asociadas
Mecanismo	¿Cómo construye la salida?	Predicción secuencial o eliminación iterativa de ruido
Tarea	¿Qué resultado necesita la persona?	Redactar, clasificar, recuperar o ilustrar

Dimensión	Pregunta pertinente	Ejemplo de respuesta
Evaluación	¿Qué propiedad debe comprobarse?	Fidelidad documental, cobertura o adecuación visual

Nota. Elaboración propia; se añaden a los cuatro ejes controles de evaluación transversales, con ejemplos independientes que no describen un único producto.

En este marco, la expresión modelo de lenguaje tampoco es sinónimo exacto de procesamiento del lenguaje natural, porque este último es un campo de tareas y técnicas más amplio; una regla que busca una palabra en un documento procesa texto, pero no por ello constituye un modelo estadístico de secuencias lingüísticas.

Modelos de lenguaje

En su formulación probabilística clásica, un modelo de lenguaje asigna probabilidades a secuencias de unidades lingüísticas; Bengio et al. (2003) plantean aprender una distribución conjunta sobre secuencias de palabras



mediante representaciones distribuidas, de manera que el modelo pueda asignar probabilidad a combinaciones no observadas literalmente en el entrenamiento.

La contribución conceptual de esa formulación no depende de que exista una ventana de conversación: permite evaluar continuaciones, comparar secuencias o muestrearlas, mientras la denominación grande se refiere a una escala relativa de tamaño y no a una propiedad que convierta automáticamente el modelo en un asistente; tamaño, interfaz y objetivo deben documentarse por separado.

Una forma autorregresiva descompone la probabilidad de una secuencia en probabilidades condicionales sucesivas; expresado sin notación especializada, el modelo estima la primera unidad, después la segunda considerando la primera y así sucesivamente, aunque el contexto disponible y la parametrización concreta determinan qué dependencias puede representar (Bengio et al., 2003).

A partir de esas probabilidades, un procedimiento de selección permite construir una continuación; esta operación necesita distinguirse del entrenamiento que



ajustó el modelo y de la aplicación que presenta la respuesta, pues una misma distribución puede utilizarse para puntuar alternativas o para producir una secuencia sin que la interfaz revele esa diferencia.

Un ejemplo mínimo de generación secuencial

Consideremos un ejemplo hipotético deliberadamente pequeño: después de la expresión «La muestra», un modelo asigna probabilidad 0.50 a «cambió», 0.30 a «permanece» y 0.20 a «desapareció»; esos valores suman uno y representan alternativas dentro del vocabulario simplificado del ejemplo, no resultados de un modelo real ni una estimación de lo que ocurrió en un laboratorio.

Si se selecciona «cambió», la distribución del siguiente paso se calcula usando esa continuación, de modo que la ruta cambia respecto de la que habría seguido «permanece»; una probabilidad de 0.40 para «ayer» condicionada a la primera elección daría a la secuencia abreviada «cambió ayer» una probabilidad conjunta condicional de 0.50 × 0.40 = 0.20 bajo estos supuestos.

El cálculo muestra dependencia secuencial, pero no acredita un cambio real en la muestra; la probabilidad corresponde a la secuencia dentro del modelo y no puede interpretarse sin más como confianza empírica en el acontecimiento, porque para sostener este último harían falta observaciones del laboratorio y un vínculo verificable entre esas observaciones y la respuesta.

Tampoco debe confundirse selección secuencial con elección arbitraria: las alternativas están condicionadas por parámetros, contexto y procedimiento de decodificación, aunque esa estructura no implica que cada afirmación producida haya sido contrastada con una fuente externa; conocer el mecanismo permite evitar tanto una explicación mágica como la reducción del resultado a una lotería sin regularidades.

Lenguaje enmascarado y representación

BERT ofrece una objeción concreta a la identificación de modelo lingüístico con chatbot: su preentrenamiento incluye la predicción de elementos enmascarados mediante contexto a ambos lados, además de una tarea de relación entre segmentos, y sus representaciones pueden adaptarse a clasificación o extracción de

respuestas sin producir una conversación abierta (Devlin et al., 2019, secciones 3 y 4).

Conviene preservar el matiz terminológico: BERT es un modelo de representaciones lingüísticas entrenado, entre otros objetivos, con modelado de lenguaje enmascarado; no es idéntico a un modelo autorregresivo de izquierda a derecha, por lo que incluir ambos bajo una acepción amplia de modelos de lenguaje exige explicar la diferencia y no afirmar que todos estiman o muestrean secuencias de la misma manera.

En una tarea hipotética de búsqueda de una fecha dentro de un párrafo, seleccionar las posiciones inicial y final de un fragmento puede resolver la necesidad sin redactar un relato; si la aplicación añade después una frase de cortesía, esa frase pertenece a la presentación del resultado y no transforma retrospectivamente la extracción en una prueba de generación abierta.

La frontera entre capacidad y uso aparece también en T5, cuyo marco transforma tareas lingüísticas en problemas de texto a texto, incluida la clasificación; producir la palabra que nombra una categoría es una salida textual generada,



pero el objetivo práctico continúa siendo elegir correctamente la categoría solicitada (Raffel et al., 2020).

De aquí se desprende una regla útil para leer demostraciones: un resultado breve no prueba ausencia de generación y un resultado lingüísticamente elaborado no prueba que toda la tarea se resolviera generando; la explicación adecuada requiere conocer el objetivo, la estructura de salida y el procedimiento concreto que produjo el resultado.

Generación fuera del lenguaje

La refutación más directa de la sinonimia procede de generadores cuyo objeto no son secuencias lingüísticas; un sistema que produce imágenes a partir de variables latentes puede ser generativo sin necesitar palabras como unidades de salida, de manera que la existencia de generación no permite deducir la existencia de un modelo de lenguaje.

Además, autorregresivo no significa lingüístico: WaveNet modela audio crudo como una secuencia de muestras acústicas y condiciona cada muestra en las anteriores, por lo que utiliza una lógica secuencial sin reducir la señal a

palabras; su trabajo original también examina usos discriminativos, lo cual vuelve a mostrar que mecanismo, modalidad y tarea son dimensiones separables (van den Oord et al., 2016).

El ejemplo acústico resulta especialmente esclarecedor porque la voz puede contener lenguaje y, sin embargo, su representación técnica incluir variaciones de señal que no corresponden a unidades léxicas; generar una onda sonora, transcribirla y determinar quién habla son operaciones distintas, aunque una aplicación reúna las tres detrás de un mismo botón.

Una narración sintética podría conservar correctamente las palabras y fallar en pronunciación, ritmo o pausas, mientras una transcripción podría reconocer las palabras y omitir información acústica; esta observación es un análisis funcional, no una clasificación de calidad de productos, y permite anticipar por qué una métrica de exactitud textual no agota la evaluación de una salida de audio.

Redes generativas adversarias

Goodfellow et al. (2014) presentan las redes generativas adversarias, o GAN, con un generador que transforma ruido en muestras y un discriminador que intenta distinguirlas de los datos de entrenamiento; ambos se ajustan mediante una interacción adversaria, mientras la generación posterior puede utilizar el generador sin repetir todo ese proceso de aprendizaje ni mantener una discusión entre dos agentes conscientes.

El calificativo adversario describe funciones de optimización y no intenciones: el discriminador proporciona una señal sobre la diferencia entre distribuciones y el generador recibe una presión de entrenamiento para modificar sus salidas; el artículo original advierte dificultades de coordinación y de diversidad, por lo que el equilibrio teórico no debe confundirse con una garantía práctica para cualquier red o conjunto de datos (Goodfellow et al., 2014, secciones 3, 4 y 6).

Como ejemplo conceptual, supongamos un archivo autorizado de motivos geométricos para telas; un generador podría producir variaciones visuales que se ajustaran al tipo de material aprendido, pero la selección

comercial seguiría requiriendo comprobar legibilidad a escala, repetición del patrón y compatibilidad con el proceso de impresión, propiedades que no quedan resueltas por la etiqueta adversaria.

Si todas las salidas fueran casi idénticas, algunas podrían resultar atractivas y aun así incumplir la necesidad de un catálogo diverso; por ello, una colección de muestras elegidas cuidadosamente para una presentación no sustituye la inspección de una muestra completa, obtenida bajo un procedimiento definido y con registro de los resultados descartados.

Autoencodificadores variacionales

En un autoencodificador variacional, o VAE por sus siglas en inglés, se articula un modelo generativo con variables latentes y un modelo de inferencia aproximada; Kingma y Welling (2014) desarrollan una forma de optimizar una cota variacional mediante reparametrización, lo que permite entrenar conjuntamente componentes de codificación e interpretación probabilística de los datos.

De manera esquemática, el codificador aproxima una distribución latente condicionada a una observación y el



decodificador describe cómo producir observaciones a partir de variables latentes; el objetivo combina ajuste de reconstrucción y regularización de esa distribución, por lo que no basta con tomar cualquier compresor de imágenes, llamarlo VAE y asumir que muestrear números arbitrarios producirá resultados adecuados (Kingma & Welling, 2014, secciones 2 y 3).

La palabra latente designa una variable que no se observa directamente en la representación original; no significa que el sistema haya identificado necesariamente conceptos humanos transparentes, de modo que una coordenada interna no puede etiquetarse como elegancia, intención o significado sin pruebas que sostengan esa interpretación.

Un caso hipotético de exploración de siluetas ayuda a distinguir reconstruir de generar: comprimir una silueta recibida y aproximarla de nuevo permite evaluar cuánto se perdió, mientras muestrear una representación para proponer otra silueta plantea una pregunta diferente sobre plausibilidad y diversidad; ninguno de esos procedimientos demuestra que la propuesta pueda fabricarse sin revisar medidas y materiales.

Difusión y eliminación de ruido

Ho et al. (2020) estudian modelos probabilísticos de difusión que aprenden transiciones capaces de revertir un proceso de incorporación gradual de ruido; en generación se parte de una muestra ruidosa y se realizan pasos sucesivos de transformación, con parámetros aprendidos y una formulación probabilística específica, hasta obtener una muestra en el espacio de datos.

La explicación de eliminar ruido es pedagógicamente útil si no sugiere que una fotografía completa estaba escondida dentro del ruido esperando ser descubierta; el resultado se construye mediante el modelo y el procedimiento de muestreo, mientras el proceso directo usado para formular el entrenamiento no es una búsqueda de una imagen real preexistente (Ho et al., 2020, secciones 2 y 3).

En consecuencia, una imagen de una plaza inexistente puede ser coherente con la instrucción sin representar un lugar fotografiado; la evaluación cambia si esa misma salida se presenta como reconstrucción de una plaza real, porque entonces la disposición de edificios, distancias y elementos visibles exige fuentes que la apariencia visual no puede proporcionar.



El número de pasos constituye una decisión técnica, pero no una ley uniforme de calidad: Nichol y Dhariwal (2021) muestran que modificaciones en los modelos de difusión permiten reducir evaluaciones durante el muestreo con cambios pequeños de calidad en sus experimentos; ese resultado respalda estudiar intercambios entre costo y calidad, no prometer que cualquier reducción conservará el desempeño de cualquier implementación.

Sistemas híbridos y multimodales

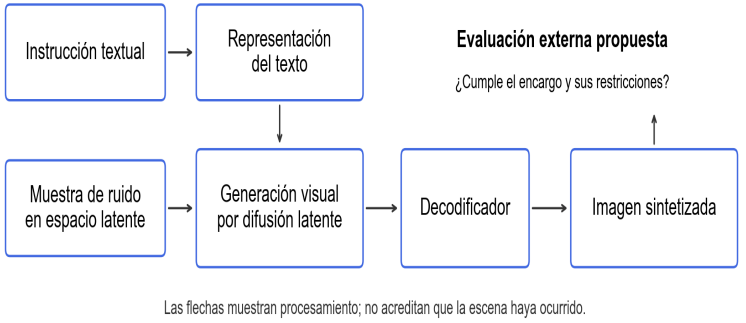
Las familias descritas no forman compartimentos excluyentes: Rombach et al. (2022) combinan una etapa de autoencodificación con difusión en un espacio latente y mecanismos de condicionamiento, incluido texto; esa composición permite evitar parte del costo de operar directamente en el espacio de píxeles, conforme a los experimentos y supuestos de su propuesta.

En una descripción simplificada de difusión latente condicionada por texto, la instrucción se representa mediante un componente textual, la generación trabaja sobre una representación visual latente y un decodificador produce la imagen; esta cadena explica por qué reconocer

un componente lingüístico no autoriza atribuirle, por sí solo, todos los pasos de síntesis visual (Rombach et al., 2022, sección 3).

La Figura 1 representa esa separación funcional sin reproducir una arquitectura comercial particular; las flechas indican dependencias de procesamiento y no relaciones causales con el mundo representado, de modo que el esquema no sugiere que la salida sea una fotografía ni que el componente textual verifique la realidad de la escena.

Figura 1. *Funciones diferentes dentro de una generación visual condicionada*



Nota. Diagrama conceptual de elaboración propia a partir de la separación entre condicionamiento, difusión latente y decodificación expuesta por Rombach et al. (2022); la evaluación externa es una propuesta de este artículo y no un componente atribuido a su arquitectura.

También puede generarse una imagen de manera autorregresiva: Ramesh et al. (2021) presentan un sistema que modela conjuntamente tokens de texto e imagen en una secuencia, mediante un transformer y una representación discreta de las imágenes; este caso impide establecer la regla simplista de que todo texto es autorregresivo y toda imagen procede de difusión.

El ejemplo exige conservar la fecha y la identidad de la investigación: describe el trabajo publicado en 2021 y no determina la arquitectura de cualquier producto posterior que reutilice un nombre relacionado; una marca puede permanecer mientras cambian componentes, entrenamiento o mecanismos, de modo que las afirmaciones técnicas deben referirse a una versión documentada.



Multimodal no significa necesariamente generativo

CLIP utiliza pares de imágenes y textos para aprender representaciones mediante un objetivo contrastivo, que favorece la asociación de pares correspondientes frente a otros pares del lote; permite comparar imágenes con descripciones y apoyar clasificación o recuperación, sin ser por ese solo objetivo un generador de imágenes ni un redactor de conversaciones (Radford et al., 2021, sección 2).

Su importancia para la distinción es doble: conecta modalidades sin tener que sintetizar una nueva imagen y demuestra que utilizar lenguaje como supervisión no equivale a generar lenguaje; un componente de ese tipo puede integrarse después en un sistema generativo, pero la propiedad del sistema completo no debe atribuirse indiscriminadamente a cada componente.

Supongamos que una biblioteca visual recibe la consulta «puente de piedra al atardecer» y devuelve archivos existentes; aunque la consulta esté escrita en lenguaje natural y la búsqueda use representaciones multimodales, el criterio principal es recuperar imágenes pertinentes del archivo, mientras una herramienta que produce un puente

nuevo necesita controles sobre la imagen sintetizada y su identificación como tal.

Diferencias entre familias

La Tabla 2 resume diferencias de mecanismo sin establecer una clasificación universal de superioridad; las descripciones corresponden a formulaciones representativas de los trabajos citados y admiten combinaciones, variantes y mejoras posteriores, por lo que deben usarse para formular preguntas técnicas y no para descartar automáticamente una familia.

Tabla 2. *Familias representativas y preguntas que orientan su evaluación*

Familia	Operación característica	Pregunta que no resuelve la etiqueta
Autorregresiva	Construye secuencias mediante distribuciones condicionales sucesivas	¿La secuencia satisface hechos y restricciones de la tarea?

Familia	Operación característica	Pregunta que no resuelve la etiqueta
GAN	Aprende generación mediante una interacción con un discriminador	¿Las muestras cubren suficiente diversidad?
VAE	Articula variables latentes, inferencia aproximada y decodificación	¿Qué se pierde al representar y reconstruir los datos?
Difusión	Aprende transformaciones que permiten muestrear mediante eliminación de ruido	¿Qué calidad y costo produce el procedimiento elegido?
Híbrida	Combina representaciones y mecanismos especializados	¿Dónde aparecen errores y cómo se propagan?

Nota. Síntesis de elaboración propia basada en Bengio et al. (2003), Goodfellow et al. (2014), Kingma & Welling

(2014), Ho et al. (2020) y Rombach et al. (2022); las preguntas de evaluación son propuestas del artículo.

Una comparación responsable necesita fijar el conjunto de datos, las condiciones de uso y el presupuesto relevante; si una implementación emplea otra resolución, otro procedimiento de muestreo o un conjunto distinto de ejemplos, una diferencia de puntuación no puede atribuirse automáticamente al nombre de la familia.

También hay que distinguir costo de entrenamiento, costo por resultado y costo de revisión: un sistema podría producir una salida con rapidez y exigir correcciones extensas, mientras otro requeriría más cómputo y reduciría parte del trabajo posterior; sin medir esas etapas, el precio de una llamada o el tiempo de generación ofrecen una imagen incompleta de utilidad.

No existe una garantía de que las categorías actuales permanezcan nítidas: los sistemas híbridos ya combinan componentes y objetivos, pero esa mezcla no vuelve inútiles las distinciones; al contrario, exige aplicarlas por componente, etapa y tarea, en lugar de buscar una sola palabra que explique toda la aplicación.



Calidad y comprobación

Una imagen puede ser nítida y omitir un objeto solicitado; un texto puede ser correcto en gramática y atribuir a una fuente una conclusión que no contiene; una voz puede resultar natural y pronunciar mal una cifra, por lo que la evaluación necesita criterios ligados a la función del contenido y no solamente a su apariencia.

Theis et al. (2016) muestran que diferentes criterios para evaluar modelos generativos, como la verosimilitud y la calidad visual de muestras, no mantienen una equivalencia general en espacios de alta dimensión; su análisis impide deducir desempeño en una aplicación a partir de una sola propiedad conveniente, y fundamenta evaluar directamente el uso previsto.

Aquí distinguimos plausibilidad, fidelidad, exactitud y diversidad: plausibilidad significa que el resultado parece compatible con una clase de contenidos; fidelidad significa que conserva información o restricciones de una fuente; exactitud supone correspondencia con un criterio verificable, y diversidad alude a la variedad relevante entre



resultados, sin convertir cada variación superficial en una contribución nueva.

La originalidad requiere otra pregunta: qué aporta la obra frente a antecedentes y cómo se relaciona con materiales previos; producir una secuencia distinta de caracteres no demuestra originalidad académica, del mismo modo que un parecido estilístico no basta, por sí solo, para concluir una infracción jurídica, asunto que necesita hechos y análisis propios.

Métricas útiles con alcances limitados

Heusel et al. (2017) introducen la distancia de Fréchet Inception, conocida como FID, para comparar estadísticas de representaciones de imágenes reales y generadas; se trata de una evaluación de distribuciones en un espacio de características y no de un certificado individual de verdad, procedencia o cumplimiento de cada instrucción textual.

Si una colección consigue una puntuación favorable, todavía podría contener imágenes inadecuadas para un encargo particular; esta es una consecuencia del nivel de agregación de la evaluación, no una afirmación de que la métrica carezca de utilidad, pues puede ayudar a



comparar configuraciones cuando el procedimiento, las muestras y las condiciones están controlados.

Kynkäänniemi et al. (2019) proponen examinar por separado precisión y cobertura de distribuciones generadas, con medidas basadas en representaciones de las muestras; la separación permite estudiar el intercambio entre fidelidad y variedad que una cifra agregada puede ocultar, aunque esos términos técnicos no deben confundirse con exactitud factual de un texto ni con cobertura documental de una investigación.

Para el lenguaje, un resumen necesita conservar lo sustentado por su fuente: Maynez et al. (2020) documentan contenido no fiel en sistemas de resumen abstractivo y muestran limitaciones de métricas habituales para capturar ese problema; el resultado se refiere a los sistemas, datos y evaluación de su estudio, sin establecer una tasa universal de errores para todos los modelos posteriores.

El perfil de riesgos de Autio et al. (2024) identifica la confabulación como producción de contenido falso o erróneo presentado de forma convincente y recomienda

verificar fuentes y citas; esa orientación institucional es una guía de gestión y no un experimento que demuestre el desempeño de una herramienta específica, distinción necesaria al combinar publicaciones técnicas y documentos de política.

Un promedio puede ocultar el error que importa

Imaginemos, como cálculo ilustrativo, un lote de 100 resúmenes: 90 cumplen el criterio de fidelidad definido y 10 contienen errores; el porcentaje global sería 90 %, pero esa cifra no distingue si los 10 errores son erratas menores o alteraciones de fechas, cantidades y obligaciones, de modo que no basta para decidir si el sistema es aceptable.

Si los 10 errores se concentran en los 20 documentos con tablas complejas, el desempeño de ese subgrupo es 10 de 20, equivalente a 50 %, aunque el promedio general conserve el mismo valor; este ejemplo no describe un producto real, sino el efecto matemático de combinar grupos con dificultades distintas.

Una evaluación útil debe definir antes qué constituye un error, cuál es su gravedad y qué grupos tienen sentido para



la tarea; también debe conservar denominadores, casos excluidos y procedimiento de revisión, pues seleccionar únicamente los resultados bien formados convierte una prueba de aceptación en una exhibición de éxitos.

Caso de referencias generadas

El siguiente escenario es hipotético: un equipo editorial necesita una explicación de dos páginas sobre un tema técnico y solicita cinco referencias; la herramienta entrega un texto fluido y una bibliografía con apariencia académica, pero la necesidad real comprende dos tareas distintas, redactar una explicación y recuperar evidencia que sostenga cada afirmación relevante.

La revisión inicial identifica cinco entradas: tres corresponden a publicaciones existentes y dos no pueden localizarse con los datos proporcionados; entre las tres existentes, una aborda un asunto cercano sin sostener la proposición atribuida, por lo que contar enlaces o verificar que una revista existe no alcanza para aprobar el aparato bibliográfico.



En este ejemplo, dos de cinco referencias cumplen tanto existencia como pertinencia para la atribución concreta, equivalente a 40 %; no es una medida de exactitud global del texto, ya que varias afirmaciones pueden depender de una misma fuente y otras carecer de cita, de modo que la unidad de revisión debe ser también la proposición y su relación con la evidencia.

Un procedimiento adecuado separaría afirmaciones verificables, interpretaciones y ejemplos; para cada afirmación decisiva, localizaría la publicación original, comprobaría autoría y versión, examinaría el pasaje correspondiente y restringiría la redacción cuando la fuente solo respaldara una conclusión más estrecha, sin llenar lagunas mediante obras inventadas ni citas genéricas.

La clasificación tecnológica ayuda a comprender el fallo: un modelo puede producir correctamente la estructura formal de una referencia sin haber consultado la obra descrita, mientras una herramienta de búsqueda puede encontrar una obra auténtica sin demostrar que respalda la frase; combinar ambos componentes solo mejora el



trabajo si la correspondencia se verifica en el resultado final.

La respuesta tampoco debe reducirse a pedir al mismo sistema que asegure que todo es verdadero, porque una afirmación adicional de seguridad no incorpora evidencia nueva; conviene solicitar localizadores y comprobarlos, conservar incertidumbres y distinguir una búsqueda fallida de la demostración de que una obra no existe.

La aceptación editorial podría exigir cero referencias inexistentes y revisión de todas las afirmaciones decisivas, junto con una lista de asuntos que permanecen sin resolver; este umbral es una propuesta para el escenario y no una garantía general, ya que la revisión humana también puede equivocarse si carece de tiempo, conocimiento o acceso a las fuentes.

Caso de reconstrucción visual

Un segundo escenario hipotético presenta dos encargos visualmente parecidos: ilustrar una ciudad imaginaria para un ensayo y reconstruir el aspecto de una calle histórica; ambos pueden solicitarse mediante texto, pero el primero



permite inventar rasgos coherentes con el concepto, mientras el segundo exige conservar evidencia sobre edificios, ubicación, fecha y elementos conocidos.

Para la ilustración conceptual, la evaluación puede revisar composición, legibilidad, relación con el argumento y ausencia de símbolos que cambien su sentido; identificarla como ilustración generada evita confundirla con un registro documental, aunque esa etiqueta no excuse errores si el dibujo pretende explicar un mecanismo técnico concreto.

En la reconstrucción, una ventana, un letrero o una fachada plausibles podrían ser adiciones no documentadas; el responsable tendría que distinguir rasgos respaldados por fotografías o planos, inferencias justificadas y elementos desconocidos, además de decidir si esos elementos se omiten, se señalan como inciertos o se muestran en versiones alternativas claramente identificadas.

No debe interpretarse el aspecto fotográfico como mayor fuerza probatoria: una imagen de alta resolución puede reproducir con gran detalle una hipótesis equivocada, mientras un esquema sencillo puede expresar con más



honestidad qué se conoce; la elección del formato tiene consecuencias sobre lo que el lector cree estar observando.

Para un diagrama de arquitectura de software, la prioridad sería conservar nodos, conexiones y rótulos verificables; si el generador visual añade una flecha atractiva pero incorrecta, el resultado incumple el encargo, de modo que construir el esquema con herramientas de diagramación y relaciones explícitas puede ser una solución más adecuada que pedir una imagen libre.

La revisión puede descomponer el encargo en restricciones observables: número de elementos, orientación, relación espacial, texto exacto y ausencia de conexiones adicionales; esta descomposición permite discutir errores concretos y evita que la impresión general de belleza absorba toda la calificación de un documento que pretende enseñar cómo funciona algo.

Caso de búsqueda multimodal

Un tercer escenario hipotético parte de un archivo de 500 imágenes catalogadas y una necesidad muy precisa:

encontrar fotografías existentes de un tipo de maquinaria; generar imágenes nuevas sería contraproducente porque la persona necesita documentos del archivo y sus datos de procedencia, aunque un sistema generativo pudiera producir representaciones convincentes del mismo objeto.

Una solución de recuperación compara la consulta con las representaciones del archivo y devuelve candidatos; el resultado debe evaluarse contra la pertinencia definida para el trabajo, sin asumir que compartir palabras o rasgos visuales implique corresponder al equipo buscado, especialmente cuando varias máquinas tienen apariencia semejante.

Supongamos que el archivo contiene 20 imágenes pertinentes, el sistema devuelve 10 y 8 de esas 10 son pertinentes; la precisión de recuperación sería 8 dividido entre 10, equivalente a 80 %, y la exhaustividad sería 8 dividido entre 20, equivalente a 40 %, según una anotación de referencia que también debería revisarse.

Estos cálculos muestran una diferencia de uso: si la persona necesita una sola ilustración de ejemplo, ocho resultados pertinentes pueden ser suficientes, pero si debe



inventariar todas las fotografías de una inspección, dejar sin recuperar 12 imágenes pertinentes representa una omisión relevante; el mismo sistema y la misma consulta pueden servir para una finalidad y fallar para otra.

La precisión y exhaustividad de este ejemplo se calculan sobre documentos recuperados, mientras las medidas de distribuciones generadas mencionadas antes se construyen con otra metodología; compartir términos no autoriza intercambiar denominadores ni interpretar una puntuación visual como porcentaje de fotografías pertinentes encontradas.

Un componente generativo podría añadirse para describir resultados o redactar una respuesta resumida, pero esa incorporación abriría una nueva superficie de error: la descripción podría atribuir detalles que no aparecen en las imágenes o mezclar metadatos de archivos distintos; la utilidad del componente debe comprobarse por separado y después en el recorrido completo.

Novedad y memorización pueden coexistir

La afirmación de que todo contenido generado es nuevo y nunca reproduce material previo no resiste la evidencia: Carlini et al. (2021) muestran extracción de secuencias presentes en datos de entrenamiento de GPT-2 mediante consultas al modelo, lo que establece la posibilidad de memorización extraíble bajo las condiciones estudiadas, sin demostrar que todas las respuestas sean copias.

En el ámbito visual, Carlini et al. (2023) documentan extracción de ejemplos de entrenamiento de modelos de difusión; este hallazgo contradice una garantía absoluta de ausencia de reproducción y obliga a considerar riesgos de procedencia y privacidad, aunque tampoco permite inferir, sin pruebas, que cualquier imagen individual corresponda a un ejemplo memorizado.

La relación entre generalización y memorización no se resuelve mediante una etiqueta comercial: un sistema puede producir combinaciones no observadas literalmente y conservar información específica de algunos ejemplos, de manera que verificar originalidad, privacidad o cercanía con antecedentes exige procedimientos adecuados a esas preguntas y acceso suficiente a los materiales pertinentes.

Para el trabajo editorial, conviene evitar dos conclusiones simétricamente excesivas: declarar que todo resultado es independiente del entrenamiento y declarar que todo resultado es una reproducción identificable; ambas exceden lo que demuestra observar una salida, mientras los estudios de extracción aportan evidencia concreta de un riesgo y no una teoría jurídica completa de cada caso.

Evaluación según la tarea

La Tabla 3 propone controles diferenciados para aplicaciones posibles; su propósito es convertir las distinciones conceptuales en decisiones observables, sin presentar una lista de verificación como sustituto de una evaluación especializada ni suponer que todas las tareas requieren el mismo nivel de revisión.

Tabla 3. *Controles propuestos para tareas con resultados diferentes*

Tarea	Comprobación principal	Evidencia que conviene conservar
Resumen documental	Cada afirmación conserva el sentido de la fuente	Documento, pasaje y correcciones
Clasificación textual	La categoría coincide con el criterio definido	Casos etiquetados y errores por grupo
Ilustración conceptual	La composición comunica el concepto sin engañar	Encargo, versión y revisión editorial
Reconstrucción visual	Los detalles están respaldados o marcados como inciertos	Fuentes visuales y decisiones de reconstrucción
Recuperación de imágenes	Los archivos devueltos son pertinentes y suficientes	Consulta, archivo y anotación de referencia
Síntesis de audio	Se conservan palabras, pronunciación y uso previsto	Texto de referencia y revisión de la señal

Nota. Propuesta metodológica de elaboración propia; los controles se ajustan a cada contexto y no constituyen certificación de un sistema.

El primer paso es formular el resultado aceptable antes de elegir el modelo: redactar un borrador, recuperar un documento o reproducir fielmente una cifra son necesidades distintas, y la palabra generar no define por sí misma cuál debe satisfacerse; esta decisión inicial evita evaluar con criterios cómodos un producto que no resuelve el problema.

El segundo paso es identificar qué operaciones realiza realmente la aplicación y cuáles delega a otros componentes; si un sistema combina recuperación, clasificación y redacción, conviene examinar tanto cada operación como los puntos de transferencia, pues una fuente correcta puede llegar incompleta al generador o una cita terminar asociada con una afirmación distinta.

El tercer paso es preparar casos representativos y casos difíciles sin escogerlos después de observar qué salidas favorecen al sistema; para un resumen, podrían incluirse



documentos con negaciones, cifras parecidas o información contradictoria, mientras una tarea visual podría exigir relaciones espaciales precisas o rótulos breves cuya corrección sea verificable.

El cuarto paso es comparar contra una alternativa razonable: una plantilla, una búsqueda convencional, extracción de texto o trabajo manual pueden resolver parte de la necesidad con menor variabilidad; la comparación debe reconocer ventajas y limitaciones de cada opción, sin asumir que usar una técnica generativa es un objetivo en sí mismo.

El quinto paso es conservar todas las condiciones necesarias para interpretar el resultado: versión del sistema, entrada, configuración pertinente, fecha, criterios de selección y revisión aplicada; si solo se guarda la mejor salida, no puede reconstruirse cuánto costó obtenerla ni qué proporción de intentos incumplió el encargo.

El sexto paso es decidir qué errores bloquean el uso y quién puede autorizar una excepción; en una ficción, inventar un lugar puede ser adecuado, mientras en una ficha documental inventar una fecha debe impedir la

liberación, de modo que el umbral de aceptación surge del propósito y de las consecuencias, no del entusiasmo que produzca la demostración.

Objeciones y límites de la distinción

Una primera objeción sostiene que, si una herramienta resuelve el trabajo, su clasificación importa poco; el argumento es razonable para usos de bajo riesgo donde el resultado se comprueba fácilmente, pero pierde fuerza cuando se necesitan procedencia, cobertura o reproducción consistente, porque conocer los componentes ayuda a elegir pruebas y localizar fallos.

Otra objeción advierte que los sistemas integrados hacen obsoletas las categorías: una aplicación puede recibir voz, interpretar imágenes, redactar texto y producir audio; sin embargo, esa integración no elimina diferencias de función, sino que vuelve más importante especificar qué componente produce cada salida y qué propiedad se está evaluando en cada etapa.

Existe también una dificultad legítima de vocabulario: algunos usos restringen modelo de lenguaje a

distribuciones sobre secuencias completas y otros incluyen ampliamente modelos de representación entrenados mediante tareas lingüísticas; este artículo conserva esa ambigüedad explícita al tratar BERT, pues resolverla mediante una definición impuesta ocultaría parte del problema que se pretende explicar.

La revisión documental permite demostrar que la sinonimia es incorrecta mediante contraejemplos y contrastes, pero no determina qué sistema concreto será mejor para cada organización; esa decisión requiere pruebas con sus tareas, datos permitidos y restricciones reales, además de observar cambios de versión que podrían alterar resultados previos.

Tampoco se infieren comprensión humana, intención o conciencia a partir de estas capacidades: los mecanismos descritos permiten explicar operaciones y evaluar productos, mientras atribuciones sobre experiencia subjetiva exigen argumentos distintos; evitar antropomorfismos no obliga a negar la utilidad de una herramienta ni a reducir todas sus funciones a una sola metáfora.

La incertidumbre relevante es, por tanto, específica: cuánto se conserva de una fuente, cuánta diversidad se cubre, cuánto cuesta revisar una salida y cómo cambian esos resultados al modificar el contexto; responder esas preguntas mediante observaciones es más productivo que discutir si una denominación amplia garantiza inteligencia o creatividad en cualquier situación.



Conclusión

IA generativa y modelos de lenguaje no son sinónimos porque describen dimensiones diferentes del trabajo computacional: la primera expresión destaca producción de contenido o una forma de modelado, mientras la segunda se refiere al lenguaje como objeto de representación y predicción; sus relaciones se comprenden mejor al separar modalidad, objetivo, mecanismo y tarea.

Los contraejemplos son suficientes para deshacer la equivalencia: BERT permite estudiar representaciones y tareas lingüísticas sin identificarlo con un chatbot autorregresivo, CLIP conecta modalidades sin sintetizar imágenes por su objetivo contrastivo y WaveNet genera audio mediante una estructura secuencial; GAN, VAE y difusión amplían el campo de generación más allá del lenguaje y pueden combinarse en sistemas híbridos.

La consecuencia práctica es que producir contenido y acreditar su adecuación son operaciones distintas: ni la fluidez demuestra fidelidad documental, ni la nitidez demuestra realidad, ni una puntuación agregada demuestra ausencia de errores relevantes; la evaluación

debe seguir el propósito y conservar evidencia de las propiedades que efectivamente se comprobaron.

Referencias

Autio, C., Schwartz, R., Dunietz, J., Jain, S., Stanley, M., Tabassi, E., Hall, P., y Roberts, K. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* [Marco de gestión de riesgos de inteligencia artificial: Perfil de inteligencia artificial generativa] (NIST AI 600-1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.600-1

Bengio, Y., Ducharme, R., Vincent, P., y Jauvin, C. (2003). A neural probabilistic language model [Un modelo probabilístico neuronal de lenguaje]. *Journal of Machine Learning Research, 3*, 1137–1155. https://www.jmlr.org/papers/v3/bengio03a.html

Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., y Wallace, E. (2023). Extracting training data from diffusion models [Extracción de datos de entrenamiento de modelos de difusión]. En *32nd USENIX Security Symposium* (pp. 5253–5270). USENIX Association. https://www.usenix.org/conference/usenixsecurity23/presentation/carlini

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., y Raffel, C. (2021). Extracting training data from large language models [Extracción de datos de entrenamiento de modelos grandes de lenguaje]. En *30th USENIX Security Symposium* (pp. 2633–2650). USENIX Association. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

Devlin, J., Chang, M.-W., Lee, K., y Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding [BERT: Preentrenamiento de transformers bidireccionales profundos para la comprensión del lenguaje]. En J. Burstein, C. Doran, y T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., y Bengio, Y. (2014). Generative adversarial nets [Redes generativas adversarias]. *Advances in Neural Information Processing Systems, 27*, 2672–2680. https://proceedings.neurips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., y Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local Nash equilibrium [Las GAN entrenadas mediante una regla de actualización de dos escalas temporales convergen a un equilibrio local de Nash]. *Advances in Neural Information Processing Systems, 30*, 6626–6637. https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html

Ho, J., Jain, A. N., y Abbeel, P. (2020). Denoising diffusion probabilistic models [Modelos probabilísticos de difusión con eliminación de ruido]. *Advances in Neural Information Processing Systems, 33*, 6840–6851. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html

Kingma, D. P., y Welling, M. (2014). *Auto-encoding variational Bayes* [Inferencia bayesiana variacional mediante autoencodificación] (versión 10) [Prepublicación]. arXiv. https://doi.org/10.48550/arXiv.1312.6114

Kynkäänniemi, T., Karras, T., Laine, S., Lehtinen, J., y Aila, T. (2019). Improved precision and recall metric for assessing generative models [Métrica mejorada de precisión y cobertura para evaluar modelos generativos]. *Advances in Neural Information Processing Systems, 32*, 3927–3936. https://proceedings.neurips.cc/paper/2019/hash/0234c510bc6d908b28c70ff313743079-Abstract.html

Maynez, J., Narayan, S., Bohnet, B., y McDonald, R. (2020). On faithfulness and factuality in abstractive summarization [Sobre fidelidad y factualidad en el resumen abstractivo]. En D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

Ng, A. Y., & Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive Bayes [Sobre clasificadores discriminativos frente a generativos: Comparación de regresión logística y Bayes ingenuo]. *Advances in Neural Information Processing Systems, 14*, 841–848. https://proceedings.neurips.cc/paper/2001/hash/7b7a53e239400a13bd6be6c91c4f6c4e-Abstract.html

Nichol, A. Q., y Dhariwal, P. (2021). Improved denoising diffusion probabilistic models [Modelos probabilísticos mejorados de difusión con eliminación de ruido]. *Proceedings of Machine Learning Research, 139*, 8162–8171. https://proceedings.mlr.press/v139/nichol21a.html

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., y Sutskever, I. (2021). Learning transferable visual models from natural language supervision [Aprendizaje de modelos visuales transferibles mediante supervisión en lenguaje natural]. *Proceedings of Machine Learning Research, 139*, 8748–8763. https://proceedings.mlr.press/v139/radford21a.html

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., y Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer [Exploración de los límites del aprendizaje por transferencia con un transformer unificado de texto a texto]. *Journal of Machine Learning Research, 21*(140), 1–67. https://jmlr.org/papers/v21/20-074.html

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., y Sutskever, I. (2021). Zero-shot text-to-image generation [Generación de imágenes a partir de texto sin ejemplos específicos de la tarea]. *Proceedings of Machine Learning Research, 139*, 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., y Ommer, B. (2022). High-resolution image synthesis with latent diffusion models [Síntesis de imágenes de alta resolución con modelos de difusión latente]. En *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695). IEEE. https://doi.org/10.1109/CVPR52688.2022.01042

Theis, L., van den Oord, A., & Bethge, M. (2016). *A note on the evaluation of generative models* [Una nota sobre la evaluación de modelos generativos] (versión 3) [Prepublicación]. arXiv. https://doi.org/10.48550/arXiv.1511.01844

van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., y Kavukcuoglu, K. (2016). *WaveNet: A generative model for raw audio* [WaveNet: Un modelo generativo para audio crudo] (versión 2) [Prepublicación]. arXiv. https://doi.org/10.48550/arXiv.1609.03499

