

Listado de componentes de IA

Se presenta una lista de tecnologías, componentes y definiciones relacionadas a la IA para comprender mejor su funcionamiento y alcances desarrollado por Andressen Horowitz. <https://a16z.com/>.

Parte con una introducción general a los modelos *de difusión latente* y *de transformadores*, que están alimentando la ola actual de IA. A continuación, profundizamos en los recursos de aprendizaje técnico; guías prácticas para construir con grandes modelos de lenguaje (LLM); y análisis del mercado de IA. Finalmente, incluimos una lista de referencia de resultados de investigaciones históricas, comenzando con "La atención es todo lo que necesita", el documento de 2017 de Google que presentó al mundo los modelos de transformadores y marcó el comienzo de la era de la IA generativa.

Una introducción general

Estos artículos no requieren conocimientos especializados y pueden ayudarlo a ponerse al día rápidamente en las partes más importantes de la ola moderna de IA.

- [Software 2.0](#) : Andrej Karpathy fue uno de los primeros en explicar claramente (en 2017!) por qué la nueva ola de IA es realmente importante. Su argumento es que la IA es una forma nueva y poderosa de programar computadoras. Como los LLM han mejorado rápidamente, esta tesis ha demostrado ser profética y brinda un buen modelo mental de cómo puede progresar el mercado de la IA.
- [Estado de GPT](#) : también de Karpathy, esta es una explicación muy accesible de cómo funcionan los modelos ChatGPT / GPT en general, cómo usarlos y qué direcciones puede tomar I+D.
- [¿Qué está haciendo ChatGPT... y por qué funciona?](#) : El científico informático y empresario Stephen Wolfram brinda una explicación larga pero muy amena, desde los primeros principios, de cómo funcionan los modelos modernos de IA. Sigue la línea de tiempo desde las primeras redes neuronales hasta los LLM y ChatGPT de hoy.
- [Entendiendo el Modelo GPT](#) : Esta publicación de Dale Markowitz es una respuesta más corta y directa a la pregunta "¿qué es un LLM y cómo funciona?" Esta es una excelente manera de facilitar el tema y desarrollar la intuición para la tecnología. Se escribió sobre GPT-3, pero aún se aplica a los modelos más nuevos.
- [Cómo funciona Stable Diffusion](#) : Esta es la visión por computadora análoga a la última publicación. Chris McCormick da una explicación sencilla de cómo funciona Stable Diffusion y desarrolla la intuición en torno a los modelos de texto a imagen.

en general. Para una introducción aún *más suave*, echa un vistazo a este [cómico](#) de r/StableDiffusion.

- [Glosario de IA de a16z](#): Definiciones de los términos y tecnologías clave con los que se encontrará cuando lea sobre las técnicas modernas de IA.

Aprendizaje fundamental: redes neuronales, retropropagación e incrustaciones

Estos recursos brindan una comprensión básica de las ideas fundamentales en el aprendizaje automático y la IA, desde los conceptos básicos del aprendizaje profundo hasta los cursos de nivel universitario de expertos en IA.

Explicadores

- [Aprendizaje profundo en pocas palabras: conceptos básicos](#) : esta serie de cuatro partes de Nvidia recorre los conceptos básicos del aprendizaje profundo tal como se practicó en 2015, y es un buen recurso para cualquiera que esté aprendiendo sobre IA.
- [Aprendizaje profundo práctico para programadores](#) : curso completo y gratuito sobre los fundamentos de la IA, explicados a través de ejemplos prácticos y código.
- [Explicación de Word2vec](#) : introducción fácil a incrustaciones y tokens, que son componentes básicos de LLM (y todos los modelos de lenguaje).
- [Sí, debe comprender backprop](#) : una publicación más detallada sobre la propagación hacia atrás si desea comprender los detalles. Si quiere aún más, pruebe la [conferencia Stanford CS231n](#) en Youtube.

Cursos

- [Stanford CS229](#) : Introducción al aprendizaje automático con Andrew Ng, que cubre los fundamentos del aprendizaje automático.
 - [Stanford CS224N](#) : PNL con aprendizaje profundo con Chris Manning, que cubre los conceptos básicos de PNL a través de la primera generación de LLM.
-

Análisis profundo de la tecnología: comprender los transformadores y los modelos grandes

Hay innumerables recursos, algunos mejores que otros, que intentan explicar cómo funcionan los LLM. Estos son algunos de nuestros favoritos, dirigidos a una amplia gama de lectores/espectadores.

Explicadores

- [El transformador ilustrado](#) : una descripción más técnica de la arquitectura del transformador por Jay Alammar.
- [El transformador anotado](#) : publicación detallada si desea comprender los transformadores a nivel de código fuente. Requiere algún conocimiento de PyTorch.
- [Construyamos GPT: desde cero, en código, explicado](#) : Para los ingenieros, Karpathy hace un tutorial en video sobre cómo construir un modelo GPT.
- [La difusión estable ilustrada](#) : Introducción a los modelos de difusión latente, el tipo más común de modelo de IA generativa para imágenes.
- [RLHF: aprendizaje reforzado a partir de la retroalimentación humana](#) : Chip Huyen explica RLHF, que puede hacer que los LLM se comporten de manera más predecible y amigable para los humanos. Este es uno de los aspectos más importantes pero menos entendidos de sistemas como ChatGPT.
- [Aprendizaje por refuerzo a partir de la retroalimentación humana](#) : el científico informático y cofundador de OpenAI, John Schulman, profundiza en esta gran charla sobre el estado actual, el progreso y las limitaciones de los LLM con RLHF.

Cursos

- [Stanford CS25](#) : Transformers United, un seminario en línea sobre Transformers.
- [Stanford CS324](#) : modelos de lenguaje grande con Percy Liang, Tatsuru Hashimoto y Chris Re, que cubren una amplia gama de aspectos técnicos y no técnicos de los LLM.

Referencia y comentario

- [Aprendizaje predictivo. NIPS 2016](#) : En esta charla inicial, Yann LeCun presenta un caso sólido para el aprendizaje no supervisado como un elemento crítico de las arquitecturas modelo de IA a escala. Pase al [minuto 19:20](#) para ver la famosa analogía del pastel, que sigue siendo uno de los mejores modelos mentales para la IA moderna.

- [IA para la conducción autónoma en Tesla](#) : otra charla clásica de Karpathy, esta vez sobre el motor de recopilación de datos de Tesla. Comenzar [alas 8:35](#) es una de las mejores diatribas de la IA de todos los tiempos, que explica por qué los problemas de cola larga (en este caso, la detección de señales de alto) son tan difíciles.
- [La hipótesis del escalado](#) : uno de los aspectos más sorprendentes de los LLM es que el escalado (agregar más datos y cómputo) sigue aumentando la precisión. GPT-3 fue el primer modelo en demostrar esto claramente, y la publicación de Gwern hace un gran trabajo al explicar la intuición detrás de esto.
- [Implicaciones descabelladas de Chinchilla](#) : nominalmente una explicación del importante artículo de Chinchilla (ver más abajo), esta publicación llega al corazón de la gran pregunta en la escala de LLM: ¿nos estamos quedando sin datos? Esto se basa en la publicación anterior y brinda una visión actualizada de las leyes de escala.
- [Una encuesta de grandes modelos lingüísticos](#) : desglose completo de los LLM actuales, incluido el cronograma de desarrollo, el tamaño, las estrategias de capacitación, los datos de capacitación, el hardware y más.
- [Chispas de inteligencia artificial general: Primeros experimentos con GPT-4](#) : Análisis temprano de Microsoft Research sobre las capacidades de GPT-4, el LLM más avanzado actual, en relación con la inteligencia humana.
- [La revolución de la IA: cómo Auto-GPT desencadena una nueva era de automatización y creatividad](#) : una introducción a Auto-GPT y los agentes de IA en general. Esta tecnología es muy temprana, pero es importante comprenderla: utiliza el acceso a Internet y subtarefas autogeneradas para resolver problemas u objetivos específicos y complejos.
- [El efecto Waluigi](#) : nominalmente una explicación del "efecto Waluigi" (es decir, por qué surgen "alter egos" en el comportamiento de LLM), pero interesante principalmente por su inmersión profunda en la teoría de las indicaciones de LLM.

Guías prácticas para construir con LLM

Está surgiendo una nueva pila de aplicaciones con LLM en el núcleo. Si bien todavía no hay mucha educación formal disponible sobre este tema, sacamos algunos de los recursos más útiles que hemos encontrado.

Referencia

- [Cree un bot de soporte de GitHub con GPT3, LangChain y Python](#) : una de las primeras explicaciones públicas de la pila de aplicaciones LLM moderna. Algunos de los consejos aquí están anticuados, pero en muchos sentidos dieron inicio a la adopción y experimentación generalizadas de nuevas aplicaciones de IA.

- [Creación de aplicaciones LLM para producción](#) : Chip Huyen analiza muchos de los desafíos clave en la creación de aplicaciones LLM, cómo abordarlos y qué tipos de casos de uso tienen más sentido.
- [Guía de ingeniería de Prompt](#): para cualquier persona que escriba avisos de LLM, incluidos los desarrolladores de aplicaciones, esta es la guía más completa, con ejemplos específicos para un puñado de modelos populares. Para un tratamiento más ligero y conversacional, pruebe [la guía de ingeniería rápida de Brex](#) .
- [Inyección inmediata: ¿Qué es lo peor que puede pasar?](#) La inyección inmediata es una vulnerabilidad de seguridad potencialmente grave que acecha para las aplicaciones LLM, sin una solución perfecta todavía. Simon Willison da la descripción definitiva del problema en esta publicación. Casi todo lo que Simon escribe sobre IA es sobresaliente.
- [Libro de recetas de OpenAI](#) : para desarrolladores, esta es la colección definitiva de guías y ejemplos de código para trabajar con la API de OpenAI. Se actualiza continuamente con nuevos ejemplos de código.
- [Centro de aprendizaje Pinecone](#) : muchas aplicaciones LLM se basan en un paradigma de búsqueda vectorial. El centro de aprendizaje de Pinecone, a pesar de ser contenido de marca del proveedor, ofrece algunas de las instrucciones más útiles sobre cómo desarrollar este patrón.
- [Documentos de LangChain](#) : como la capa de orquestación predeterminada para las aplicaciones LLM, LangChain se conecta a casi todas las demás piezas de la pila. Por lo tanto, sus documentos son una referencia real para la pila completa y cómo encajan las piezas.
- [Arquitecturas emergentes para aplicaciones LLM](#) : una mirada a la arquitectura y las herramientas que los desarrolladores están utilizando para crear aplicaciones LLM con aprendizaje en contexto, desde API hasta bases de datos vectoriales.
- [Primeros pasos con AI Stack para JavaScript](#) : una pila de inicio rápido para desarrolladores que desean experimentar con IA generativa sin pasar días investigando los componentes correctos. [El repositorio de GitHub está aquí](#).

Cursos

- [LLM Bootcamp](#) : un curso práctico para crear aplicaciones basadas en LLM con Charles Frye, Sergey Karayev y Josh Tobin.
- [Hugging Face Transformers](#) : guía para usar LLM de código abierto en la biblioteca de transformadores Hugging Face.

Puntos de referencia de LLM

- [Chatbot Arena](#) : un sistema de clasificación al estilo Elo de LLM populares, dirigido por un equipo de UC Berkeley. Los usuarios también pueden participar comparando modelos cara a cara.

- [Open LLM Leaderboard](#) : una clasificación de Hugging Face, que compara los LLM de código abierto en una colección de puntos de referencia y tareas estándar.

Análisis de mercado

Todos nos hemos maravillado con lo que puede producir la IA generativa, pero todavía hay muchas preguntas sobre *lo que significa todo esto* . ¿Qué productos y empresas sobrevivirán y prosperarán? ¿Qué les pasa a los artistas? ¿Cómo deberían usarlo las empresas? ¿Cómo afectará literalmente a los puestos de trabajo ya la sociedad en general? Aquí hay algunos intentos de responder a estas preguntas.

a16z pensando

- [¿Quién es el propietario de la plataforma de IA generativa?](#) : Nuestra evaluación principal de dónde se acumula y podría acumular valor en las capas de infraestructura, modelo y aplicación de la IA generativa.
- [Navegando por el alto costo de la computación de IA](#) : un desglose detallado de por qué los modelos de IA generativa requieren tantos recursos informáticos y cómo pensar en adquirir esos recursos (es decir, las GPU correctas en la cantidad correcta, al costo correcto) en un alto costo. mercado de la demanda.
- [El arte no está muerto, solo es generado por máquinas](#) : una mirada a cómo los modelos de IA pudieron remodelar los campos creativos, que a menudo se supone que son el último obstáculo contra la automatización, mucho más rápido que campos como el desarrollo de software.
- [La revolución de la IA generativa en los juegos](#) : un análisis en profundidad de nuestro equipo de juegos sobre cómo la capacidad de crear fácilmente gráficos altamente detallados cambiará la forma en que funcionan los diseñadores de juegos, los estudios y todo el mercado. [Este artículo de seguimiento](#) de nuestro equipo de Juegos analiza específicamente la llegada del contenido generado por IA frente al contenido generado por el usuario.
- [Para aplicaciones de IA generativa B2B, ¿menos es más?](#) : Una predicción de cómo evolucionarán los LLM en el mundo de las aplicaciones empresariales B2B, centrada en la idea de que, en última instancia, resumir información será más valioso que producir texto.
- [Los servicios financieros adoptarán la IA generativa más rápido de lo que piensa](#) : un argumento de que la industria de servicios financieros está lista para usar la IA generativa para experiencias personalizadas del consumidor, operaciones rentables, mejor cumplimiento, mejor gestión de riesgos y pronósticos e informes dinámicos.

- [IA generativa: la próxima plataforma de consumo](#) : una mirada a las oportunidades para que la IA generativa impacte en el mercado de consumo en una variedad de sectores, desde la terapia hasta el comercio electrónico.
- [Para marcar una diferencia real en el cuidado de la salud, la IA deberá aprender como nosotros](#) : la IA está lista para cambiar irrevocablemente la forma en que buscamos prevenir y tratar enfermedades. Sin embargo, para transformar realmente el descubrimiento de fármacos en la prestación de atención, debemos invertir en la creación de un ecosistema de IA "especializadas", que aprenden como lo hacen nuestros mejores médicos y desarrolladores de fármacos en la actualidad.
- [La nueva revolución industrial: Bio x AI](#) : La próxima revolución industrial en la historia humana será la biología impulsada por inteligencia artificial.

Otras perspectivas

- [Sobre las oportunidades y los riesgos de los modelos básicos](#) : documento general de Stanford sobre los modelos básicos. Largo y obstinado, pero esto dio forma al término.
- [Informe sobre el estado de la IA](#) : un resumen anual de todo lo que sucede en la IA, incluidos los avances tecnológicos, el desarrollo de la industria, la política/regulación, las implicaciones económicas, la seguridad y las predicciones para el futuro.
- [Los GPT son GPT: una mirada temprana al potencial de impacto en el mercado laboral de los modelos de lenguaje grande](#) : este artículo de investigadores de OpenAI, OpenResearch y la Universidad de Pensilvania predice que "alrededor del 80% de la fuerza laboral de EE. UU. podría tener al menos el 10% de sus tareas laborales se ven afectadas por la introducción de los LLM, mientras que aproximadamente el 19 % de los trabajadores pueden ver afectadas al menos el 50 % de sus tareas".
- [Medicina profunda: cómo la inteligencia artificial puede hacer que la atención médica vuelva a ser humana](#) : el Dr. Eric Topol revela cómo la inteligencia artificial tiene el potencial de liberar a los médicos de las tareas que consumen mucho tiempo y que interfieren con la conexión humana. Se restablece la relación médico-paciente. ([podcast de a16z](#))

Resultados de investigaciones históricas

La mayoría de los sorprendentes productos de IA que vemos hoy en día son el resultado de investigaciones no menos sorprendentes, realizadas por expertos dentro de grandes empresas y universidades líderes. Últimamente, también hemos visto un trabajo impresionante de personas y de la comunidad de código abierto que llevan proyectos populares hacia nuevas

direcciones, por ejemplo, mediante la creación de agentes automatizados o la migración de modelos a espacios de hardware más pequeños.

Aquí hay una colección de muchos de estos documentos y proyectos, para personas que realmente quieren profundizar en la IA generativa. (Para trabajos y proyectos de investigación, también hemos incluido enlaces a las publicaciones de blog o sitios web adjuntos, cuando estén disponibles, que tienden a explicar las cosas en un nivel superior. Y hemos incluido años de publicación originales para que pueda realizar un seguimiento de la investigación fundamental a lo largo del tiempo. .)

Grandes modelos de lenguaje

Nuevos modelos

- [La atención es todo lo que necesitas](#) (2017): el trabajo de transformador original y el documento de investigación de Google Brain que lo inició todo. ([entrada de blog](#))
- [BERT: entrenamiento previo de transformadores bidireccionales profundos para la comprensión del idioma](#) (2018): uno de los primeros LLM disponibles públicamente, con muchas variantes todavía en uso en la actualidad. ([entrada de blog](#))
- [Mejorar la comprensión del lenguaje mediante el entrenamiento previo generativo](#) (2018): el primer artículo de OpenAI que cubre la arquitectura GPT, que se ha convertido en la ruta de desarrollo dominante en los LLM. ([entrada de blog](#))
- [Language models are few-shot learners](#) (2020): el artículo de OpenAI que describe GPT-3 y la arquitectura de solo decodificador de los LLM modernos.
- [Entrenamiento de modelos de lenguaje para seguir instrucciones con retroalimentación humana](#) (2022): artículo de OpenAI que explica InstructGPT, que utiliza humanos en el ciclo para entrenar modelos y, por lo tanto, seguir mejor las instrucciones en las indicaciones. Este fue uno de los desbloques clave que hizo que los LLM fueran accesibles para los consumidores (por ejemplo, a través de ChatGPT). ([entrada de blog](#))
- [LaMDA: modelos de lenguaje para aplicaciones de diálogo](#) (2022): un modelo de Google diseñado específicamente para un diálogo fluido entre un humano y un chatbot en una amplia variedad de temas. ([entrada de blog](#))
- [PaLM: Modelado de lenguaje escalable con vías](#) (2022): PaLM, de Google, utilizó un nuevo sistema para entrenar LLM en miles de chips y demostró mejoras mayores de lo esperado para ciertas tareas a medida que aumentaba el tamaño del modelo. ([entrada de blog](#)). Consulte también el [informe técnico de PaLM-2](#).
- [OPT: Modelos de lenguaje de transformadores preentrenados abiertos](#) (2022): OPT es uno de los LLM de código abierto de mayor rendimiento. El lanzamiento de este modelo de 175 mil millones de parámetros viene con código y se entrenó en conjuntos de datos disponibles públicamente. ([entrada de blog](#))
- [Entrenamiento de modelos de lenguaje grande con optimización de](#) cómputo (2022): The Chinchilla paper. Argumenta que la mayoría de los modelos tienen

datos limitados, no computación limitada, y cambió el consenso sobre la escala LLM. ([entrada de blog](#))

- [Informe técnico de GPT-4](#) (2023): ¡El último y mejor documento de OpenAI, conocido principalmente por lo poco que revela! ([entrada de blog](#)). La Tarjeta del sistema GPT-4 arroja algo de luz sobre cómo OpenAI trata las alucinaciones, la privacidad, la seguridad y otros temas.
- [LLaMA: Modelos de lenguaje base abiertos y eficientes](#) (2023): El modelo de Meta que (casi) inició una revolución LLM de código abierto. Competitivo con muchos de los mejores modelos de código cerrado, pero solo abierto a investigadores con una licencia restringida. ([entrada de blog](#))
- [Alpaca: un modelo de seguimiento de instrucciones fuerte y replicable](#) (2023): fuera de Stanford, este modelo demuestra el poder del ajuste de instrucciones, especialmente en modelos de código abierto más pequeños, en comparación con la escala pura.

Mejoras en el modelo (por ejemplo, ajuste fino, recuperación, atención)

- [Aprendizaje por refuerzo profundo a partir de las preferencias humanas](#) (2017): investigación sobre el aprendizaje por refuerzo en contextos de juegos y robótica, que resultó ser una herramienta fantástica para los LLM.
- [Generación de recuperación aumentada para tareas de PNL intensivas en conocimiento](#) (2020): Desarrollado por Facebook, RAG es una de las dos principales vías de investigación para mejorar la precisión de LLM a través de la recuperación de información. ([entrada de blog](#))
- [Mejorar los modelos de lenguaje mediante la recuperación de billones de tokens](#) (2021): RETRO, para "Retrieval Enhanced TRANSformers", es otro enfoque, este de DeepMind, para mejorar la precisión de LLM al acceder a información no incluida en sus datos de capacitación. ([entrada de blog](#))
- [LoRA: Adaptación de bajo rango de modelos de lenguaje grande](#) (2021): esta investigación de Microsoft introdujo una alternativa más eficiente al ajuste para capacitar a los LLM en nuevos datos. Ahora se ha convertido en un estándar para el ajuste fino de la comunidad, especialmente para los modelos de imágenes.
- [AI constitucional \(2022\)](#) : el equipo de Anthropic presenta el concepto de aprendizaje reforzado a partir de AI Feedback (RLAIF). La idea principal es que podemos desarrollar un asistente de IA inofensivo con la supervisión de otras IA.
- [FlashAttention: atención exacta rápida y eficiente en memoria con IO-awareness](#) (2022): esta investigación de Stanford abrió la puerta a modelos de última generación para comprender secuencias de texto más largas (e imágenes de mayor resolución) sin un entrenamiento exorbitante tiempos y costos. ([entrada de blog](#))
- [Hipopótamos hambrientos y hambrientos: hacia el modelado del lenguaje con modelos de espacio de estado](#) (2022): nuevamente de Stanford, este artículo describe una de las principales alternativas a la atención en el modelado del

lenguaje. Este es un camino prometedor hacia una mejor escalabilidad y eficiencia en el entrenamiento. ([entrada de blog](#))

Modelos de generación de imágenes

- [Aprendizaje de modelos visuales transferibles a partir de la supervisión del lenguaje natural](#) (2021): documento que presenta un modelo base, CLIP, que vincula descripciones textuales a imágenes. Uno de los primeros usos efectivos a gran escala de modelos básicos en visión por computadora. ([entrada de blog](#))
- [Generación de texto a imagen de tiro cero](#) (2021): este es el documento que presentó DALL-E, un modelo que combina el CLIP y GPT-3 antes mencionados para generar automáticamente imágenes basadas en indicaciones de texto. Su sucesor, DALL-E 2, iniciaría el auge de la IA generativa basada en imágenes en 2022. ([entrada de blog](#))
- [Síntesis de imágenes de alta resolución con modelos de difusión latente](#) (2021): el artículo que describía la difusión estable (después del lanzamiento y el crecimiento explosivo del código abierto).
- [Modelos fotorrealistas de difusión de texto a imagen con comprensión profunda del lenguaje](#) (2022): Imagen fue la incursión de Google en la generación de imágenes de IA. Más de un año después de su anuncio, el modelo aún no se ha lanzado públicamente a la fecha de publicación de este artículo. ([sitio web](#))
- [DreamBooth: ajuste fino de modelos de difusión de texto a imagen para la generación basada en sujetos](#) (2022): DreamBooth es un sistema, desarrollado en Google, para entrenar modelos para reconocer temas enviados por usuarios y aplicarlos al contexto de un mensaje (por ejemplo, [USUARIO] sonriendo a la Torre Eiffel). ([sitio web](#))
- [Agregar control condicional a los modelos de difusión de texto a imagen](#) (2023): este documento de Stanford presenta ControlNet, una herramienta ahora muy popular para ejercer un control detallado sobre la generación de imágenes con modelos de difusión latente.

Agentes

- [Un camino hacia la inteligencia artificial autónoma](#) (2022): una propuesta del líder de Meta AI y profesor de la Universidad de Nueva York, Yann LeCun, sobre cómo construir agentes autónomos e inteligentes que realmente entiendan el mundo que los rodea.
- [ReAct: sinergia del razonamiento y la actuación en modelos lingüísticos](#) (2022): un proyecto de Princeton y Google para probar y mejorar las habilidades de razonamiento y planificación de los LLM. ([entrada de blog](#))
- [Agentes generativos: simulacros interactivos del comportamiento humano](#) (2023): los investigadores de Stanford y Google utilizaron LLM para impulsar a los agentes,

en un entorno similar a "Los Sims", cuyas interacciones son emergentes en lugar de programadas.

- [Reflexión: un agente autónomo con memoria dinámica y autorreflexión](#) (2023): trabajo de investigadores de la Universidad Northeastern y el MIT sobre cómo enseñar a los LLM a resolver problemas de manera más confiable aprendiendo de sus errores y experiencias pasadas.
- [Toolformer: los modelos de lenguaje pueden aprender a usar herramientas](#) (2023): este proyecto de Meta capacitó a los LLM para usar herramientas externas (API, en este caso, apuntando a cosas como motores de búsqueda y calculadoras) para mejorar la precisión sin aumentar el tamaño del modelo.
- [Auto-GPT: un experimento autónomo de GPT-4](#) : un experimento de código abierto para expandir las capacidades de GPT-4 brindándole una colección de herramientas (acceso a Internet, almacenamiento de archivos, etc.) y eligiendo cuáles usar para resolver una tarea específica.
- [BabyAGI](#) : este script de Python utiliza GPT-4 y bases de datos vectoriales (para almacenar contexto) para planificar y ejecutar una serie de tareas que resuelven un objetivo más amplio.

Otras modalidades de datos

Código de GENERACION

- [Evaluación de grandes modelos de lenguaje entrenados en código](#) (2021): Este es el trabajo de investigación de OpenAI para Codex, el modelo de generación de código detrás del producto GitHub Copilot. ([entrada de blog](#))
- [Generación de código a nivel de competencia con AlphaCode](#) (2021): esta investigación de DeepMind demuestra un modelo capaz de escribir mejor código que los programadores humanos. ([entrada de blog](#))
- [CodeGen: un modelo abierto de lenguaje grande para código con síntesis de programa de varios turnos](#) (2022): CodeGen sale del brazo de investigación de IA en Salesforce y actualmente respalda el producto Replit Ghostwriter para la generación de código. ([entrada de blog](#))

Generación de vídeo

- [Make-A-Video: generación de texto a video sin datos de texto y video](#) (2022): un modelo de Meta que crea videos cortos a partir de indicaciones de texto, pero también agrega movimiento a entradas de fotos estáticas o crea variaciones de videos existentes. ([entrada de blog](#))
- [Imagen Video: Generación de video de alta definición con modelos de difusión](#) (2022): Justo lo que parece: una versión del modelo Imagen basado en imágenes de Google optimizado para producir videos cortos a partir de indicaciones de texto. ([sitio web](#))

Biología humana y datos médicos.

- [Strategies for pre-training graph neural networks](#) (2020): esta publicación sentó las bases para métodos efectivos de pre-entrenamiento útiles para aplicaciones en el descubrimiento de fármacos, como la predicción de propiedades moleculares y la predicción de funciones de proteínas. ([entrada de blog](#))
- [Predicción mejorada de la estructura de la proteína usando potenciales del aprendizaje profundo](#) (2020): el modelo transformador centrado en la proteína de DeepMind, AlphaFold, hizo posible predecir la estructura de la proteína a partir de la secuencia, un verdadero avance que ya ha tenido implicaciones de gran alcance para comprender los procesos biológicos y desarrollar nuevos tratamientos para enfermedades. ([entrada de blog](#)) ([explicador](#))
- [Los grandes modelos de lenguaje codifican el conocimiento clínico](#) (2022): Med-PaLM es un LLM capaz de responder correctamente preguntas del estilo del examen de licencia médica de EE. UU. Desde entonces, el equipo ha publicado resultados sobre el rendimiento de Med-PaLM2, que logró una puntuación a la par de los examinados "expertos". Otros equipos han realizado experimentos similares con [ChatGPT](#) y [GPT-4](#). ([vídeo](#))

Generación de audio

- [Jukebox: un modelo generativo para la música](#) (2020): la incursión de OpenAI en la generación de música mediante transformadores, capaces de producir música, voces y letras con una formación mínima. ([entrada de blog](#))
- [AudioLM: un enfoque de modelado de lenguaje para la generación de audio](#) (2022): AudioLM es un proyecto de Google para generar múltiples tipos de audio, incluido el habla y la instrumentación. ([entrada de blog](#))
- [MusicLM: Generación de música a partir de texto](#) (2023): estado actual del arte en la generación de música basada en IA, que muestra una mayor calidad y coherencia que los intentos anteriores. ([entrada de blog](#))

Generación de imágenes multidimensionales

- [NeRF: representación de escenas como campos de radiación neuronal para la síntesis de vistas](#) (2020): investigación de un equipo dirigido por UC-Berkeley sobre "síntesis de vistas novedosas de escenas complejas" utilizando coordenadas 5D. ([sitio web](#))
- [DreamFusion: texto a 3D mediante difusión 2D](#) (2022): trabajo de investigadores de Google y UC-Berkeley que se basa en NeRF para generar imágenes 3D a partir de entradas 2D. ([sitio web](#))