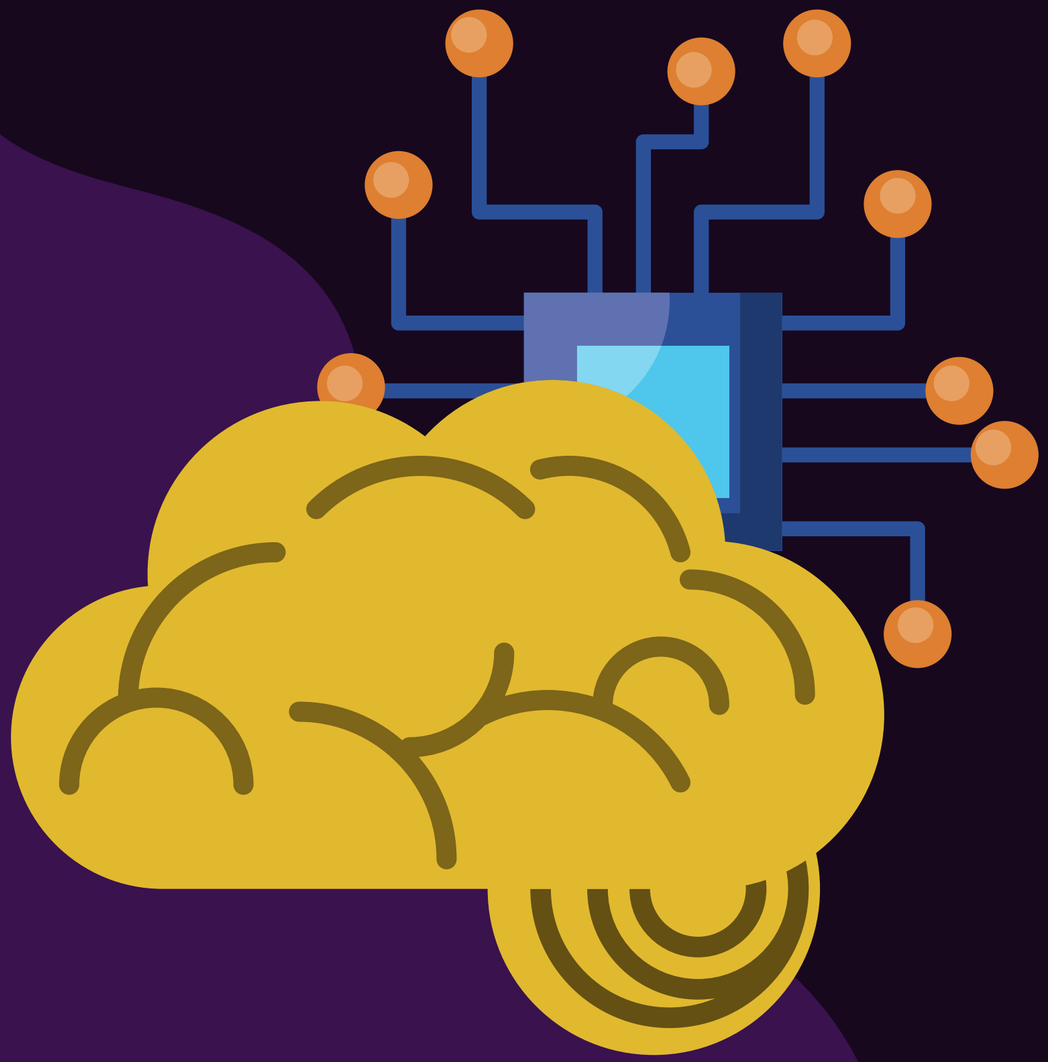
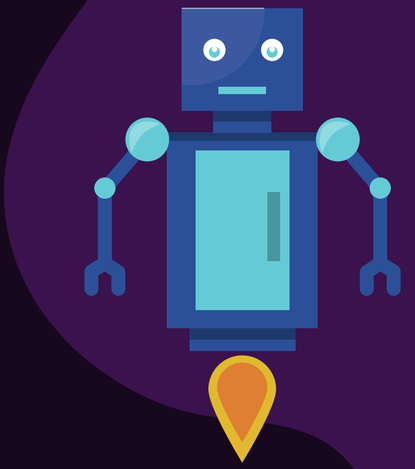


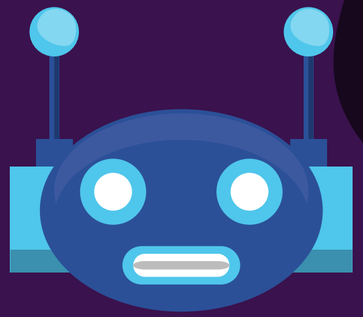
INTELIGENCIA ARTIFICIAL

Clase 14

Tutor: José Andrés Montenegro Santos
Lunes, 23 de diciembre de 2024



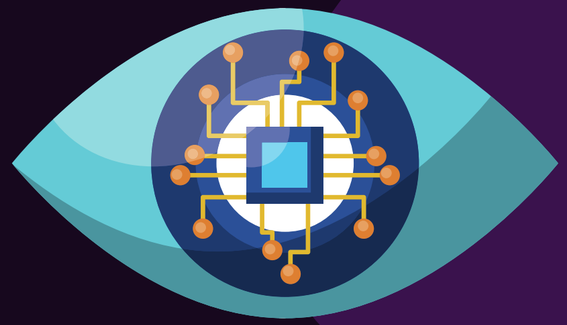
LARGE LANGUAGE MODELS (LLM)



¿Qué es?

Los modelos de lenguaje de gran tamaño (LLM) son una categoría de modelos fundacionales entrenados sobre enormes cantidades de datos que los hacen capaces de comprender y generar lenguaje natural, entre otros tipos de contenidos, para realizar una amplia gama de tareas.

Son modelos de aprendizaje profundo muy grandes que se preentrenan con grandes cantidades de datos. El transformador subyacente es un conjunto de redes neuronales que consta de un codificador y un decodificador con capacidades de autoatención. El codificador y el decodificador extraen significados de una secuencia de texto y comprenden las relaciones entre las palabras y las frases que contiene.
Los transformadores LLM son capaces de entrenarse sin supervisión, aunque una explicación más precisa es que los transformadores llevan a cabo un autoaprendizaje. Es a través de este proceso que los transformadores aprenden a entender la gramática, los idiomas y los conocimientos básicos.



¿COMO FUNCIONA?

FUNCIONAMIENTO

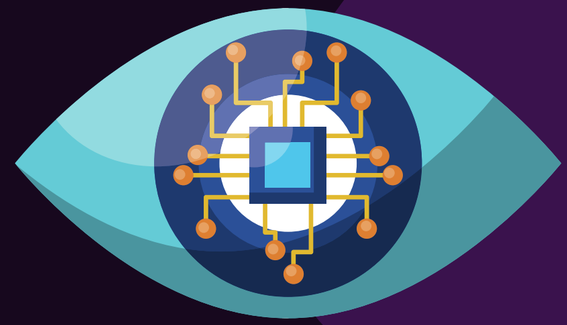
Para funcionar, los LLM aprovechan las técnicas de aprendizaje profundo y enormes cantidades de datos textuales. Estos modelos suelen basarse en una arquitectura de transformador, como el transformador generativo entrenado previamente, que se destaca en el manejo de datos secuenciales, por ejemplo, la entrada de texto. Los LLM constan de varias capas de redes neuronales, cada una con parámetros que pueden ajustarse durante el entrenamiento, y que se mejoran aún más mediante una multiplicidad de capas conocidas en su conjunto como mecanismo de atención, que se centra en partes específicas de los conjuntos de datos.

Durante el proceso de entrenamiento, estos modelos aprenden a predecir la siguiente palabra en una oración en función del contexto proporcionado por las palabras anteriores. El modelo lo hace atribuyendo una puntuación de probabilidad a la recurrencia de palabras que se han convertido en tokens, desglosadas en secuencias de caracteres más pequeñas. Luego, estos tokens se transforman en incrustaciones, que son representaciones numéricas de este contexto.

FUNCIONAMIENTO

Para garantizar la precisión, el proceso implica entrenar el LLM en un corpus masivo de texto (en miles de millones de páginas), lo que le permite aprender gramática, semántica y relaciones conceptuales mediante un aprendizaje zero-shot y autosupervisado. Una vez entrenados en estos datos, los LLM pueden generar texto al predecir de forma autónoma la siguiente palabra en función de la información que reciben y dibujar sobre los patrones y conocimientos que adquirieron. El resultado es una generación de lenguaje coherente y pertinente para el contexto que se puede aprovechar para una amplia gama de tareas de NLU y generación de contenido.

El rendimiento del modelo también se puede mejorar a través de la ingeniería rápida, el ajuste rápido, el ajuste de precisión y otras tácticas como el aprendizaje por refuerzo con retroalimentación humana (RLHF) para eliminar los sesgos, el discurso de odio y las respuestas objetivamente incorrectas conocidas como "alucinaciones" que a menudo son subproductos no deseados del entrenamiento con tantos datos no estructurados. Este es uno de los aspectos más importantes de garantizar que los LLM de nivel empresarial estén listos para su uso y no expongan a las organizaciones a responsabilidad no deseada, ni causen daños a su reputación.



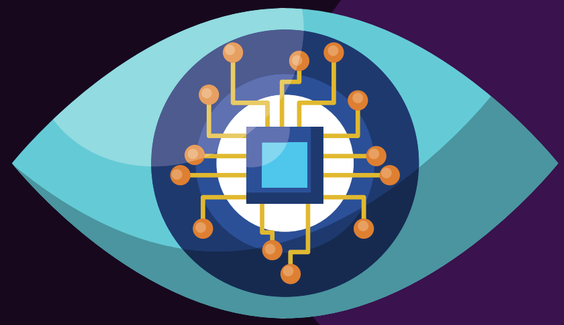
IMPORTANCIA

IMPORTANCIA

Los modelos de lenguaje de gran tamaño son increíblemente flexibles. Un modelo puede realizar tareas completamente diferentes, como responder preguntas, resumir documentos, traducir idiomas y completar oraciones. Los LLM tienen el potencial de alterar la creación de contenido y la forma en que las personas utilizan los motores de búsqueda y los asistentes virtuales.

Si bien no son perfectos, los LLM están demostrando una notable capacidad para hacer predicciones basadas en un número relativamente pequeño de indicaciones o entradas. Los LLM se pueden utilizar en la IA (inteligencia artificial) generativa para producir contenido basado en indicaciones de entrada en lenguaje humano.

Los LLM son grandes, muy grandes. Pueden considerar miles de millones de parámetros y tienen muchos usos posibles



CASOS DE USO

Casos de uso

Los LLM están redefiniendo un número cada vez mayor de procesos de negocios y han demostrado su versatilidad en una gran variedad de casos de uso y tareas en diversas industrias. Aumentan la IA conversacional en chatbots y asistentes virtuales (como IBM Watsonx Assistant y Google BARD) para mejorar las interacciones que sustentan la excelencia en la atención al cliente, proporcionando respuestas contextualizadas que imitan las interacciones con agentes humanos.

Los LLM también se destacan en la generación de contenidos, automatizando la creación de contenidos para artículos de blog, materiales de marketing o ventas y otras tareas de redacción. En la investigación y el mundo académico, ayudan a resumir y extraer información de grandes conjuntos de datos, acelerando el descubrimiento de conocimientos. Los LLM también desempeñan un papel fundamental en la traducción de idiomas, eliminando las barreras idiomáticas al proporcionar traducciones precisas y pertinentes para el contexto. Incluso pueden usarse para escribir código o "traducir" entre lenguajes de programación.

Generación de texto

Habilidades de generación de lenguaje, como escribir correos electrónicos, entradas de blog u otro contenido de formato medio o largo en respuesta a instrucciones que se pueden ajustar y pulir. Un excelente ejemplo es la generación aumentada por recuperación (RAG).

Resúmenes de contenido

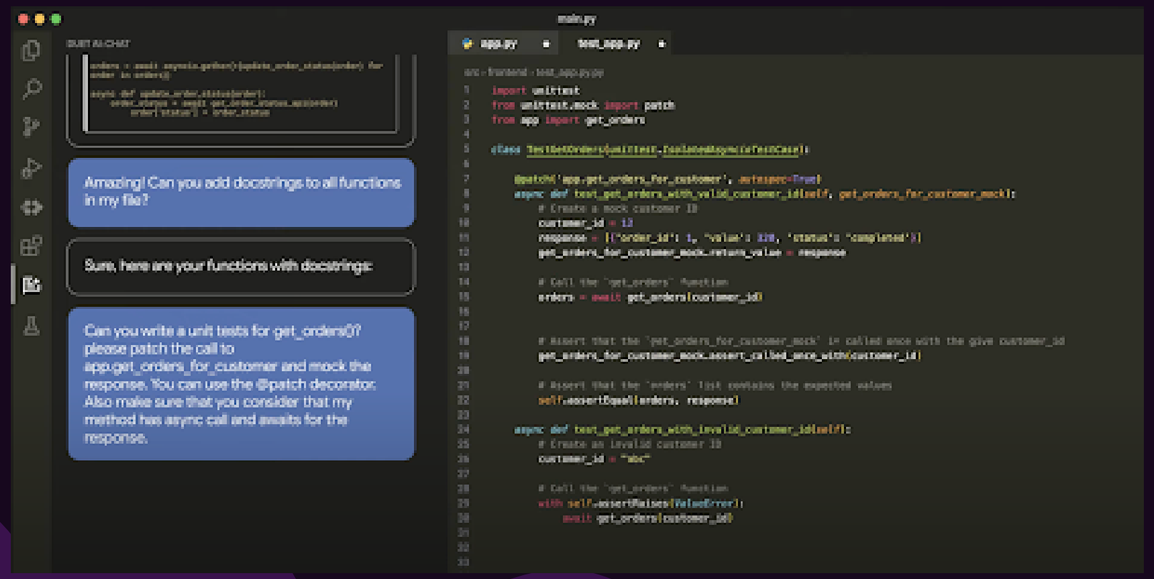
Resumir artículos largos, noticias, informes de investigación, documentación empresarial e incluso el historial de los clientes en textos completos, adaptados a la longitud del formato de salida.

Asistentes de IA

Chatbots que responden a las consultas de los clientes, realizan tareas de backend y proporcionan información detallada en lenguaje natural como parte de una solución de atención al cliente integrada y de autoservicio.

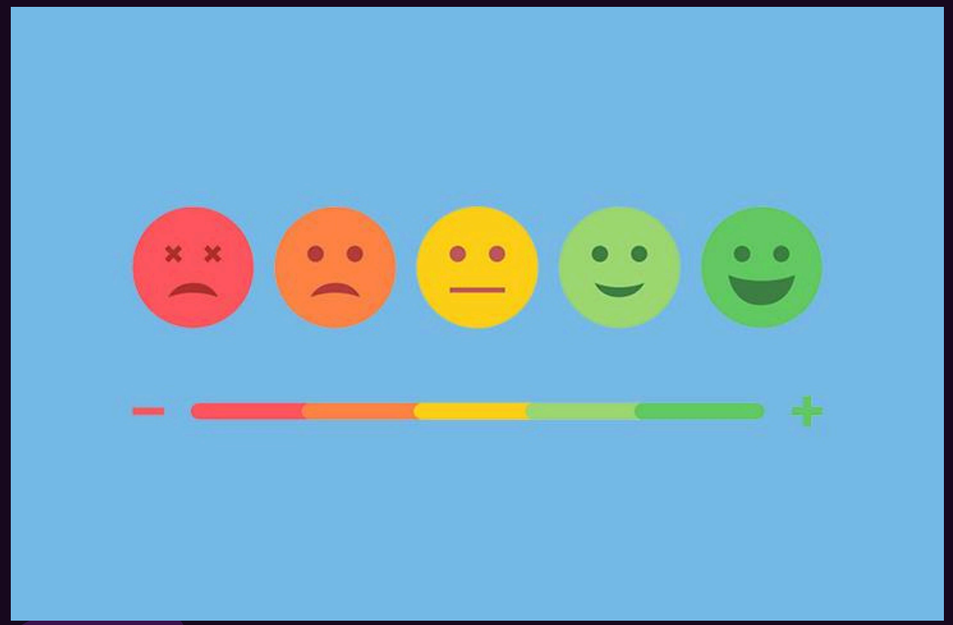
Generación de código

Ayuda a los desarrolladores a crear aplicaciones, encontrar errores en el código y detectar problemas de seguridad en varios lenguajes de programación, incluso a generar "traducciones" entre ellos.



Análisis de sentimientos

Analiza el texto para determinar el tono del cliente con el objetivo de comprender los comentarios de los clientes a escala y ayudar a gestionar la reputación de la marca.



Traducción de idiomas

Proporciona a las organizaciones una cobertura más amplia en todos los idiomas y zonas geográficas con traducciones fluidas y capacidades multilingües.