UNIVERSIDAD DE GRANADA

TRABAJO DE INVESTIGACIÓN

Aplicaciones de modelos generativos avanzados

ChatGPT

Sistemas Inteligentes para la Gestión en la Empresa

Máster Profesional en Ingeniería Informática

Curso académico 2022/2023

Autor

Ramón García Verjaga (rgarver@correo.ugr.es)

José Alberto Gómez García (modej@correo.ugr.es)

Introducción **3**
Contexto y motivación 3
Objetivos del trabajo 4
Estructura del documento 4
Fundamentos teóricos **4**
Redes neuronales y deep learning 5
Transformers 5
ChatGPT y los modelos GPT 7
Arquitectura y entrenamiento 8
Técnicas utilizadas 8
Limitaciones 11
Aplicaciones y usos reales **12**
Prompt engineering 14
Plugins 15
¿Podemos saber si un texto ha sido generado con ChatGPT? 16
Conclusiones **18**
Bibliografía **19**

Introducción

La inteligencia artificial y el procesamiento del lenguaje natural han evolucionado rápidamente en los últimos años. Uno de los protagonistas de este progreso es ChatGPT, un modelo de lenguaje creado por OpenAI. Este modelo, que se basa en la arquitectura transformer y en técnicas de aprendizaje profundo, es resultado de muchos años de trabajo e investigación en el campo de la inteligencia artificial. Este documento tiene como objetivo examinar a fondo ChatGPT, un representante destacado de lo que se conoce como inteligencia artificial de propósito general (Artificial General Intelligence, AGI). ***[OpenAI (2023) Introducing ChatGPT]***

ChatGPT ha conseguido captar rápidamente la atención de muchas personas. En tan solo dos meses, ha logrado reunir a más de 100 millones de usuarios, superando (en acogida, cantidad de usuarios captados por tiempo) incluso a algunas de las redes sociales más populares. Este modelo de lenguaje, desarrollado por OpenAI, está cambiando la manera en que nos comunicamos e interactuamos con la tecnología. Además de ser un logro en el campo de la inteligencia artificial, ChatGPT es una herramienta útil que puede proporcionar respuestas y soluciones con un nivel de comprensión y articulación lingüística muy avanzado.

Digamos que ChatGPT es como una “mente electrónica” formada por millones de conexiones neuronales, que se alimenta del conocimiento humano. Ya sea que necesitemos ayuda con tareas cotidianas, queramos aprender algo nuevo o simplemente busquemos tener una conversación, ChatGPT está listo para ayudarnos. Su impacto en la sociedad está siendo considerable, abriendo nuevas posibilidades en áreas como la educación, la salud y la industria. ChatGPT está mostrándonos que el futuro ya está aquí, que no era el tan sonado metaverso, como algunos decían, y que es más sorprendente de lo que podríamos haber imaginado.

Contexto y motivación

Dentro del escenario de la inteligencia artificial, ChatGPT se destaca como un importante avance hacia la inteligencia artificial de propósito general (AGI). La AGI se caracteriza por su capacidad para entender, aprender y aplicar sus conocimientos a una amplia gama de tareas, igual o incluso mejor que un humano. Esta AGI es considerada como el próximo hito en la evolución de la inteligencia artificial, y con ChatGPT, ya podemos vislumbrar un atisbo de su inmenso potencial. Aunque para algunos, esta visión de la AGI aún se mantiene como un espejismo, es indiscutible el camino que ChatGPT está abriendo hacia ella. ***[Some Glimpse AGI in ChatGPT. Others Call It a Mirage]***

El propósito fundamental de ChatGPT es claro: mejorar la manera en la que las máquinas interactúan con nosotros, los humanos. Busca no sólo superar las limitaciones de los modelos de lenguaje anteriores, sino también dar un paso más hacia la AGI y explorar nuevas formas en las que la IA puede enriquecer nuestra sociedad. ***[OpenAI (2023) Introducing ChatGPT]***

Objetivos del trabajo

Los objetivos de este trabajo se centran en explorar a profundidad las diversas facetas de ChatGPT. Esto implica entender su diseño, su funcionamiento y las técnicas usadas en su creación. También, se busca evaluar su impacto y usos prácticos, y cómo estos influyen en la vida diaria del consumidor promedio. Finalmente, se pretende identificar los desafíos y oportunidades futuras.

Estructura del documento

Este documento se ha estructurado para facilitar una comprensión clara y profunda de ChatGPT. Iniciamos con los fundamentos teóricos. A continuación, nos adentramos en la arquitectura de ChatGPT, las técnicas utilizadas y las limitaciones actuales. Luego exploramos las aplicaciones prácticas y cómo se utilizan técnicas como el *prompt engineering*. Finalmente, presentamos nuestras conclusiones.

Fundamentos teóricos

La inteligencia artificial (Artificial Intelligence, AI) es una rama de la informática que se centra en la creación de sistemas capaces de realizar tareas que normalmente requieren de la inteligencia humana. Estas tareas incluyen la comprensión del lenguaje natural, el reconocimiento de voz y de imágenes, el aprendizaje y la resolución de problemas.

El aprendizaje automático (Machine Learning, ML) es un subcampo de la AI que se centra en el desarrollo de algoritmos que permiten a las máquinas aprender a realizar tareas a partir de datos, en lugar de estar explícitamente programadas para hacerlo. Podríamos afirmar, de forma muy superficial, que los modelos de ML mejoran su rendimiento en una tarea determinada a medida que se exponen a más datos relacionados con esa tarea.

Dentro del ML, existen tres paradigmas principales de aprendizaje:

- **Aprendizaje supervisado**: los modelos se entrenan en un conjunto de datos etiquetado, es decir, un conjunto de datos en el que cada ejemplo viene con la respuesta correcta. Por ejemplo, si estamos entrenando un modelo para reconocer imágenes de gatos, le proporcionaríamos un conjunto de imágenes de gatos y le diríamos al modelo que todas esas imágenes son de gatos.
- **Aprendizaje no supervisado**: los modelos se entrenan en un conjunto de datos no etiquetado. El modelo debe descubrir por sí mismo patrones y relaciones en los datos. Por ejemplo, si proporcionamos al modelo un conjunto de imágenes de gatos y perros sin decirle cuáles son cuáles, el modelo tendría que aprender a distinguir entre las dos categorías por sí mismo.
- **Aprendizaje por refuerzo**: los modelos aprenden a través de la prueba y el error. Se les proporciona una función de recompensa que mide cómo de bien están realizando una

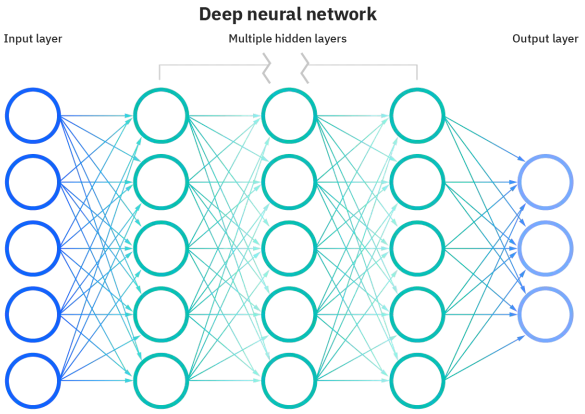
tarea, y su objetivo es maximizar esta recompensa. Este paradigma de aprendizaje es particularmente útil para enseñar a las máquinas tareas que requieren la toma de decisiones secuenciales, como jugar a un juego o mantener una conversación.

Estos conceptos fundamentales del ML son esenciales para entender cómo funcionan modelos como ChatGPT, que emplean tanto aprendizaje supervisado como por refuerzo para entrenarse y mejorar su capacidad para generar texto como el que escribiría un humano.

[AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?]

Redes neuronales y deep learning

Las redes neuronales son el pilar del aprendizaje profundo, una rama del aprendizaje automático que se ha popularizado por su eficacia en la realización de tareas complejas, especialmente aquellas relacionadas con datos no estructurados, como el texto y las imágenes.



Una red neuronal consta de una serie de "neuronas" o "nodos" organizados en capas. Cada neurona en una capa está conectada a todas las neuronas de la capa siguiente, formando una red densamente conectada. Las neuronas reciben una serie de entradas, aplican una función a esas entradas (generalmente una función no lineal) y envían la salida a las neuronas de la siguiente capa.

Cuando una red neuronal se entrena para realizar una tarea, ajusta los pesos de las conexiones entre las neuronas para minimizar el error en las predicciones del modelo. A medida que la red neuronal se expone a más datos, mejora su capacidad para realizar la tarea para la que ha sido entrenada.

[Apuntes SIGE]

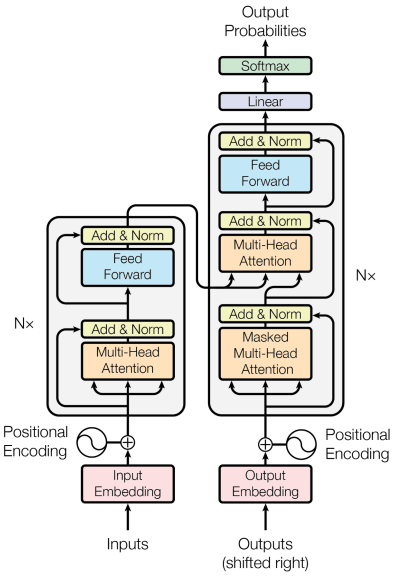
Transformers

Los transformers son un tipo de modelo de redes neuronales artificiales que se introdujo en un trabajo titulado "Attention is All You Need" por Vaswani et al. en 2017. Son especialmente útiles para tareas que implican datos secuenciales, como el procesamiento del lenguaje natural (NLP). Podríamos afirmar que son una evolución y un modelo específico dentro del amplio abanico de redes neuronales artificiales diseñado específicamente para manejar datos secuenciales. A diferencia de otros modelos de procesamiento de secuencias, como las redes neuronales

recurrentes (RNN) y las redes de memoria a largo corto plazo (LSTM), los transformers no procesan los datos secuencialmente.

[Attention Is All You Need]

Los transformers utilizan un mecanismo llamado "atención" para ponderar la relevancia de cada parte de la secuencia para cada otra parte. Esto permite manejar dependencias a largo plazo entre los elementos de una secuencia de manera más eficaz.



En lugar de tratar cada palabra con la misma importancia, los transformers prestan "atención" a las palabras más importantes. Este mecanismo de "atención" les permite concentrarse más en las palabras relevantes y menos en las palabras menos importantes. Este enfoque permite a los transformers manejar textos largos y complejos de manera más eficiente.

Este mecanismo de atención se basa en la idea de que no todas las palabras en una oración son igualmente relevantes para entender el significado de cada palabra individual. Por ejemplo, en la oración "El gato que está en el tejado es gris", las palabras "gato" y "gris" son especialmente relevantes para entender el significado de "es". Los transformers aprovechan esta idea para mejorar la calidad de las representaciones de texto que aprenden.

De forma más sencilla, a lo que nos referimos es a lo siguiente. Imaginemos que estamos leyendo un libro, no leemos cada palabra con la misma atención. Algunas palabras son más importantes para entender la historia, mientras que otras son menos importantes. Los transformers hacen algo similar cuando procesan texto.

Además, los transformers pueden procesar todas las palabras al mismo tiempo. Imaginemos que podemos leer todas las palabras de una página de un libro al mismo tiempo y entender la historia completa de una vez. Los transformers pueden hacer algo similar, lo que les permite aprender de manera más rápida y eficiente.

La arquitectura transformer ha demostrado ser muy eficaz para una amplia gama de tareas de procesamiento del lenguaje natural, y ha impulsado avances significativos en el campo.

A nivel más técnico, en lugar de depender de la recurrencia o la convolución, los transformers, como ya hemos mencionado, se basan en mecanismos de atención. La atención es una técnica que permite al modelo concentrarse en diferentes partes de la entrada en diferentes momentos,

dando más peso a las partes más relevantes para la tarea en cuestión. Esto resulta particularmente útil en tareas como la traducción automática, donde la relevancia de una palabra en la entrada puede depender de su contexto.

[The Transformer Model]

Los transformers y GPT han sido fundamentales para muchos avances recientes en el procesamiento del lenguaje natural, incluyendo modelos como ChatGPT que pueden generar texto "idéntico" al que generarían los humanos con atributos tan importantes como la creatividad y la coherencia.

[Attention Is All You Need]

ChatGPT y los modelos GPT

El GPT (Generative Pre-trained Transformer, GPT) es un modelo basado en la arquitectura transformer que ha sido preentrenado gracias a grandes cantidades de texto. GPT es un modelo generativo, lo que significa que se utiliza para generar nuevas secuencias de datos que se asemejan a los datos de entrenamiento.

ChatGPT está basado en modelos de este tipo. Ha sido diseñado para generar texto en lenguaje natural con fines conversacionales (según OpenAI, ChatGPT es muy similar a su modelo anterior, InstructGPT; no obstante, InstructGPT está diseñado para generar texto instructivo para tareas como responder preguntas o proporcionar orientación paso a paso).

Los modelos de tipo **GPT** más famosos son:

- **GPT-3** son modelos de "instrucción" que están diseñados para generar texto con una instrucción clara. No están optimizados para chat conversacional. El mejor modelo GPT-3 es *text-davinci-003*.
- **GPT-3.5** (actual ChatGPT free) se lanzaron por primera vez el 1 de marzo de 2023, nótese que ChatGPT se lanzó en noviembre de 2022. Están construidos sobre los modelos GPT-3 y optimizados para chat conversacional. Sin embargo, en la gran mayoría de los casos, son igual de buenos con instrucciones como text-davinci-003. Los resultados de GPT-3.5 pueden ser demasiado "conversadores" o "creativos" en algunos casos.
- **GPT-4** (ChatGPT plus) son la última generación de modelos de OpenAI, lanzados el 14 de marzo de 2023. Son multimodales, es decir, pueden aceptar tanto entradas de texto como de imagen; pueden resolver problemas mucho más complejos gracias a las capacidades avanzadas de razonamiento; pueden usar de dos a ocho veces más tokens en su contexto que los modelos GPT-3 y GPT-3.5; y son significativamente más caros.

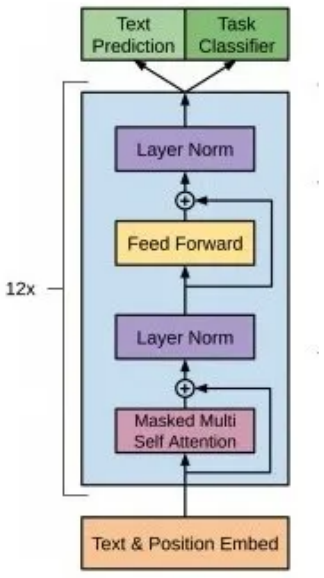
[GPT-3]

El entrenamiento de GPT consta de dos etapas principales. Primero, el modelo se preentrena en un gran corpus de texto sin etiquetas, aprendiendo a predecir la siguiente palabra en una secuencia

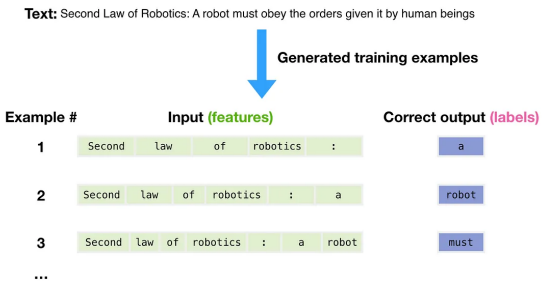
dada las palabras anteriores. Este preentrenamiento permite al modelo aprender una representación rica del lenguaje natural. En segundo lugar, el modelo se ajusta finalmente en una tarea específica (como responder preguntas o traducir texto) utilizando un conjunto de datos etiquetado más pequeño.

Arquitectura y entrenamiento

GPT-3 se ha entrenado con enormes conjuntos de datos de texto de Internet: 45 TB en total. Cuando se lanzó, era la red neuronal más grande, con 175.000 millones de parámetros. Como la mayoría de los modelos de IA, las redes neuronales son, en esencia, funciones matemáticas complejas que requieren datos numéricos como entrada. Por lo tanto, el texto de entrada se codifica primero en datos numéricos antes de introducirlo en la red.



Se utiliza un conjunto de datos con millones de tokens para generar ejemplos de entrenamiento para el modelo. Mostramos el vector de características (frase a la que le vamos a añadir la palabra) y enseñamos la palabra correcta para la salida (etiqueta). Si la palabra de salida es diferente a la que esperamos calculamos el error mostrando la salida correcta y actualizamos los parámetros del modelo, para que la próxima vez haga una predicción mejor. Este proceso se repetirá muchas veces.



Técnicas utilizadas

ChatGPT fue modificado y mejorado utilizando métodos de aprendizaje supervisado y por refuerzo, con la ayuda de un humano (Reinforcement Learning from Human Feedback, RLHF). El aprendizaje incluye 3 pasos.

Paso 1 - Modelo Supervised Fine-Tuning (SFT)

Consiste en recopilar datos de demostración para entrenar un modelo de política supervisada, denominado modelo SFT.

- Recogida de datos: se selecciona una lista de prompts y se pide a un grupo de etiquetadores humanos que escriban la respuesta de salida esperada. En el caso de ChatGPT, se han utilizado dos fuentes distintas de instrucciones:

    - algunas han sido preparadas directamente por los etiquetadores o desarrolladores,
    - y otras se han extraído de las peticiones de la API de OpenAI (es decir, de sus clientes de GPT-3).

  Como todo este proceso es lento y costoso, el resultado es un conjunto de datos relativamente pequeño y de alta calidad (de unos 12-15.000 puntos de datos, presumiblemente) que se utilizará para afinar un modelo lingüístico preentrenado.
- Elección del modelo: en lugar de perfeccionar el modelo GPT-3 original, los desarrolladores de ChatGPT optaron por un modelo preentrenado de la llamada serie GPT-3.5. Es de suponer que el modelo de referencia utilizado es el más reciente, *text-davinci-003*, un modelo GPT-3 que se ajustó, principalmente, al código de programación dando lugar a una versión GPT-3.5.

El problema es que la etapa de aprendizaje supervisado tiene un alto coste de escalabilidad, que intentaremos resolver aplicando el paso 2.

Paso 2 - Modelo de recompensa (RM)

El objetivo es aprender una función objetivo (el modelo de recompensa) directamente a partir de los datos. El propósito de esta función es dar una puntuación a las salidas del modelo SFT, proporcional a lo deseables que son estas salidas para los humanos. En la práctica, esto refleja en gran medida las preferencias específicas del grupo seleccionado de etiquetadores humanos y las directrices comunes que acordaron seguir. Al final, este proceso extraerá de los datos un sistema automático que se supone que imita las preferencias humanas.

Para lograr lo anterior:

- Se selecciona una lista de preguntas y el modelo SFT genera varios resultados (entre 4 y 9) para cada pregunta.
- Los etiquetadores clasifican los resultados de mejor a peor. El resultado es un nuevo conjunto de datos etiquetados, en el que las clasificaciones son las etiquetas. El tamaño de este conjunto de datos es aproximadamente 10 veces mayor que el conjunto de datos curados utilizado para el modelo SFT.
- Estos nuevos datos se utilizan para entrenar un modelo de recompensa (Reward Model, RM). Este modelo toma como entrada algunos de los resultados del modelo SFT y los clasifica por orden de preferencia.

Paso 3 - Fine-tuning del modelo SFT usando Proximal Policy Optimization (PPO)

El aprendizaje por refuerzo se aplica ahora para afinar la política de SFT permitiendo optimizar el modelo de recompensa. El algoritmo específico utilizado se denomina Proximal Policy Optimization (PPO) y el modelo ajustado se denomina modelo PPO.

¿Qué es PPO?

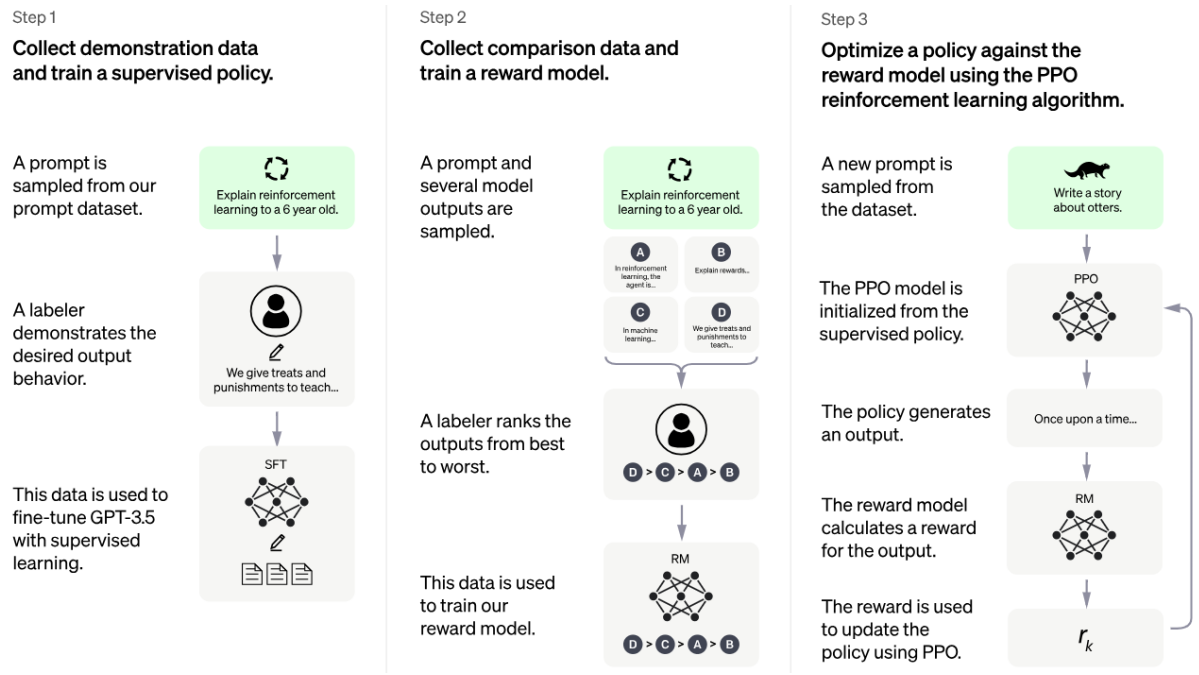
- Un algoritmo que se utiliza para entrenar agentes en el aprendizaje por refuerzo. Se denomina algoritmo "on-policy" porque aprende de la política actual y la actualiza directamente, en lugar de aprender de experiencias pasadas como en los algoritmos "off-policy" como DQN (Deep Q-Network). Esto significa que PPO adapta continuamente la política actual en función de las acciones que realiza el agente y de las recompensas que recibe.
- Utiliza un método de optimización de región de confianza para entrenar la política, lo que significa que restringe el cambio en la política para que esté dentro de una cierta distancia de la política anterior con el fin de garantizar la estabilidad.
- Utiliza una función de valor para estimar el rendimiento esperado de un estado o acción determinados. La función de valor se utiliza para calcular la función de ventaja, que representa la diferencia entre la respuesta esperada y la respuesta actual. La función de ventaja se utiliza entonces para actualizar la política comparando la acción tomada por la política actual con la acción que habría tomado la política anterior. Esto permite a PPO realizar actualizaciones más informadas de la política basándose en el valor estimado de las acciones que se están tomando.

En este paso,

- el modelo PPO se inicializa a partir del modelo SFT,
- y la función de valor se inicializa a partir del modelo de recompensa. El entorno es un entorno bandido que presenta un mensaje aleatorio y espera una respuesta al mensaje.

Dado el prompt y la respuesta, se produce una recompensa (determinada por el modelo de recompensa). Se añade una penalización KL por token del modelo SFT en cada ficha para mitigar la sobreoptimización del modelo de recompensa.

[Chat GPT and GPT 3 Detailed Architecture Study-Deep NLP Horse] [How ChatGPT actually works]

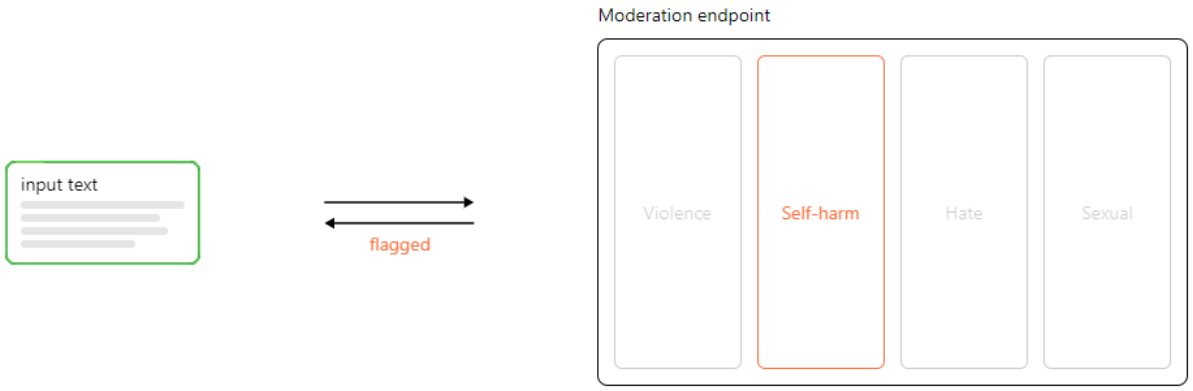


[OpenAI (2023) Introducing ChatGPT]

Limitaciones

En la sección en la que se habla sobre ChatGPT, en el propio blog de OpenAI, se hacen explícitas la siguientes limitaciones, vamos a reflexionar sobre ellas:

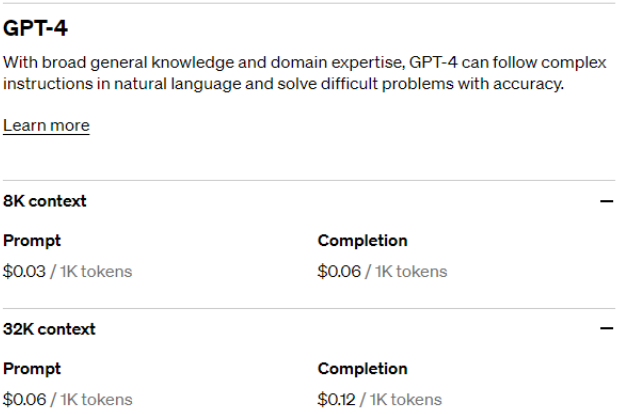
- Nos cuentan que ChatGPT a veces escribe respuestas que parecen totalmente verdaderas, pero son incorrectas o sin sentido. Esto, plantea un problema desafiante por las siguientes razones:
    - durante el entrenamiento RL (Reinforcement Learning, RL), actualmente, no existe una fuente de verdad;
    - entrenar al modelo para que sea precavido provoca que rechace preguntas que puede responder correctamente;
    - y el entrenamiento supervisado confunde al modelo porque la respuesta ideal depende de lo que el modelo sabe, en lugar de lo que sabe el humano.
- También, hacen explícito que ChatGPT es sensible a ajustes en el prompt de la entrada o al intentar usar el mismo prompt varias veces. Por ejemplo, dado un enfoque de una pregunta, el modelo puede afirmar que no sabe la respuesta, pero realizando cambios superficiales, puede responder correctamente.
- Además, explican que el modelo a menudo es excesivamente verboso y usa demasiado ciertas frases, como reafirmar que es un modelo de lenguaje entrenado por OpenAI. Estos problemas surgen de los sesgos en los datos de entrenamiento (los entrenadores prefieren respuestas más largas que parecen más completas) y otros problemas de sobreoptimización.
- Plantean que, idealmente, el modelo debería hacer preguntas aclaratorias cuando el usuario proporciona una consulta ambigua o que, por alguna razón, no s e puede llegar a comprender correctamente. Sin embargo, el modelo se centra en averiguar lo que el usuario pretende decir.
- A los puntos anteriores, se suma una de las más importantes limitaciones que ha planteado parte de la sociedad. Desde OpenAI inciden en que han hecho esfuerzos para que el modelo rechace solicitudes inapropiadas; no obstante, indican que a veces responderá a instrucciones dañinas o mostrará un comportamiento sesgado. Explican que para intentar paliar esta situación, están utilizando la que es llamada ***Moderation API*** para advertir o bloquear ciertos tipos de contenido inseguro, no obstante, no es del todo precisa, por lo que esperan obtener algunos falsos negativos y positivos. ***[New and improved content moderation tooling]***



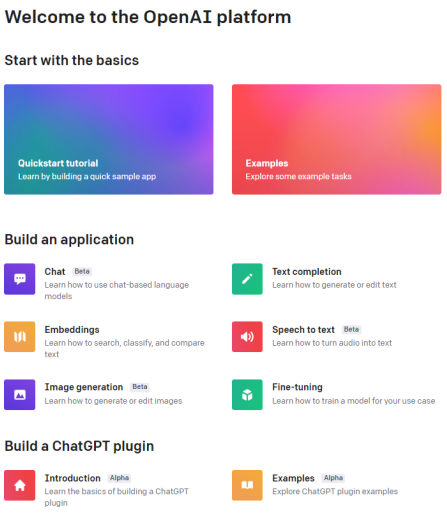
[OpenAI (2023) Introducing ChatGPT]

Aplicaciones y usos reales

Los modelos desarrollados por OpenAI, entre ellos ChatGPT, se pueden integrar en multitud de aplicaciones gracias a que se exponen a través de una API pública, con unos costes determinados en función de los tokens utilizados (***enlace a la página de precios***). ***[Pricing]***

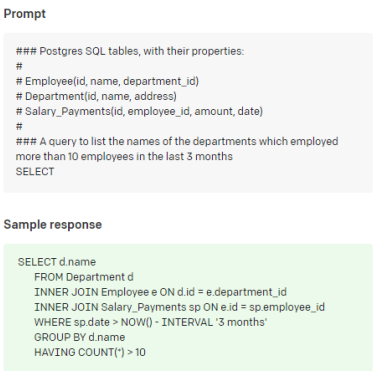


Además, desde OpenAI, ofrecen una plataforma completa para desarrolladores y personas interesadas en crear productos basados en sus tecnologías con cursos, guías documentación y ejemplos reales sobre diversos temas (***enlace a la plataforma***). ***[OpenAI platform]***



Asistentes de código

Traduce lenguaje natural a una consulta de SQL.



Generación automática de texto y resúmenes

El siguiente ejemplo convierte notas de una reunión en un breve resumen.

Prompt

Convert my short hand into a first-hand account of the meeting:

Tom: Profits up 50%
Jane: New servers are online
Kjel: Need more time to fix software
Jane: Happy to help
Parkman: Beta testing almost done

Sample response

At the meeting, Tom reported that profits had increased by 50%. Jane then mentioned that the new servers were online. Kjel mentioned that they needed more time to fix the software, and Jane offered to help. Finally, Parkman reported that the beta testing was almost done.

Otras aplicaciones relevantes

El siguiente ejemplo crea una receta a partir de una lista de ingredientes.

Prompt

Write a recipe based on these ingredients and instructions:

Frito Pie

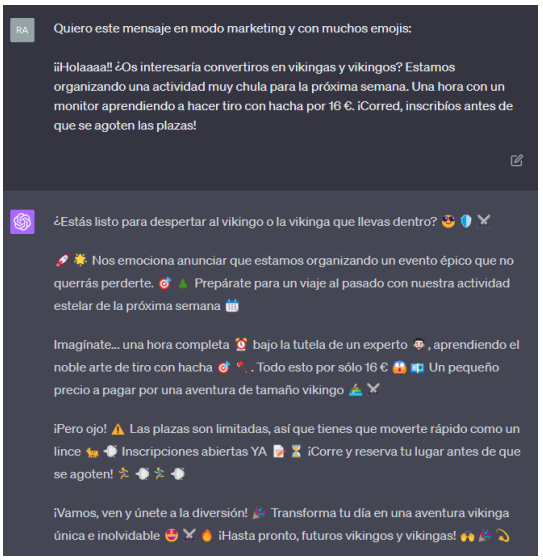
Ingredients:
Fritos
Chili
Shredded cheddar cheese
Sweet white or red onions, diced small
Sour cream

Instructions:

Sample response

1. Preheat oven to 350 degrees F.
2. Spread a layer of Fritos in the bottom of a 9x13 inch baking dish.
3. Top the Fritos with a layer of chili.
4. Sprinkle the chili with shredded cheddar cheese.
5. Sprinkle the cheese with diced onions.
6. Bake in preheated oven for 25 minutes.
7. Remove from oven and top with a dollop of sour cream.
8. Serve warm. Enjoy!

[Platform OpenAI Examples]



El siguiente ejemplo es una muestra real de una generación de un mensaje para promocionar una actividad de una asociación Erasmus. Básicamente, le decimos a ChatGPT que queremos el mensaje que le vamos a dar en modo/estilo marketing y decorado con muchos emoticonos; a partir de aquí, escribimos el mensaje con los datos clave que queremos que tenga. Si leemos el resultado, podemos apreciar que es bastante bueno en cuanto a las ideas que plasma y a las formas de hacer promoción.

Prompt engineering

Es importante hablar de este nuevo término que se acuña popularmente a partir del uso de inteligencias artificiales generativas con las que hay que interactuar a través de inputs, en muchas ocasiones de tipo textual.

Si preguntamos a ChatGPT ***¿qué es la prompt engineering?***, obtenemos la siguiente respuesta: *El término prompt engineering refiere a la ciencia y arte de diseñar, estructurar y optimizar prompts (indicaciones o estímulos) para obtener las respuestas más adecuadas de un modelo de lenguaje generativo, como GPT-4 de OpenAI.* ***[ChatGPT]***

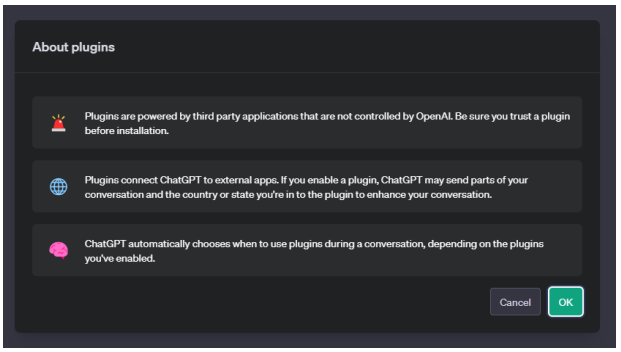
Como podemos apreciar, es una respuesta bastante acertada. El objetivo de esta disciplina es crear instrucciones precisas para generar salidas útiles y coherentes. Además, cabe destacar que, no sólo se limita a la construcción de la instrucción inicial, sino que también abarca el manejo de las respuestas del modelo y la preparación de prompts de seguimiento, para mantener un flujo de conversación natural y consistente. Esta última parte es importante, sobre todo cuando queremos construir aplicaciones. Normalmente, cuando intercambiamos mensajes con ChatGPT a través de la interfaz que nos provee, no solemos construir prompts demasiado complejos, pero cuando queremos construir una aplicación real en base al modelo, debemos hacer un gran esfuerzo para crear y controlar las interacciones *modelo - aplicación - usuario* en base a los prompts.

Como tal, aparece una nueva figura denominada *prompt engineer*. Básicamente, este nuevo rol se centra en optimizar la interacción entre los usuarios y los modelos de lenguaje a través del diseño y la mejora de los prompts. Como estos modelos aprenden de grandes cantidades de texto y no poseen una comprensión real del mundo, la calidad de la salida está fuertemente influenciada por la precisión y la claridad de los prompts. Aquí es donde entra en juego la figura del *prompt engineer*, cuyo trabajo es guiar al modelo de lenguaje hacia las respuestas más coherentes, útiles y contextuales.

Uno de los cursos gratuitos sobre esta nueva disciplina que está teniendo una mayor acogida es el impartido por los reconocidos Andrew Ng (DeepLearning.AI y Coursera) e Isa Fulford (OpenAI). El curso, titulado ***ChatGPT Prompt Engineering for Developers*** describe cómo funcionan los LLM, proporciona las mejores prácticas para la prompt engineering y muestra cómo las API de LLM se pueden utilizar en aplicaciones para una variedad de tareas. ***[ChatGPT Prompt Engineering for Developers]***

Plugins

Hablando en términos sencillos, los plugins son mejoras para ChatGPT. Es decir, piezas de software que nos ayudan a obtener información nueva, personalizada o detallada que no está presente en los datos de entrenamiento de este modelo. Además, nos permiten interactuar y realizar funciones sobre servicios externos. Cuando en nuestra cuenta (con ***suscripción plus***) vamos a activar los plugins, se nos muestra el popup informativo que vemos en la imagen. Como podemos apreciar, son tres los avisos que recibimos:



- El primero, que nos dice que nos debemos asegurar de que los plugins son fiables ya que no son controlados por OpenAI;
- el segundo, que nos explica que los plugins conectan ChatGPT a aplicaciones externas, por lo que podrían enviar datos de la conversación o la situación geográfica para mejorar la conversación;
- el tercero y último, que nos dice que en función de los plugins que tengamos activos, utilizará unos u otros automáticamente en las conversaciones si así lo considera necesario.

[ChatGPT plugins]

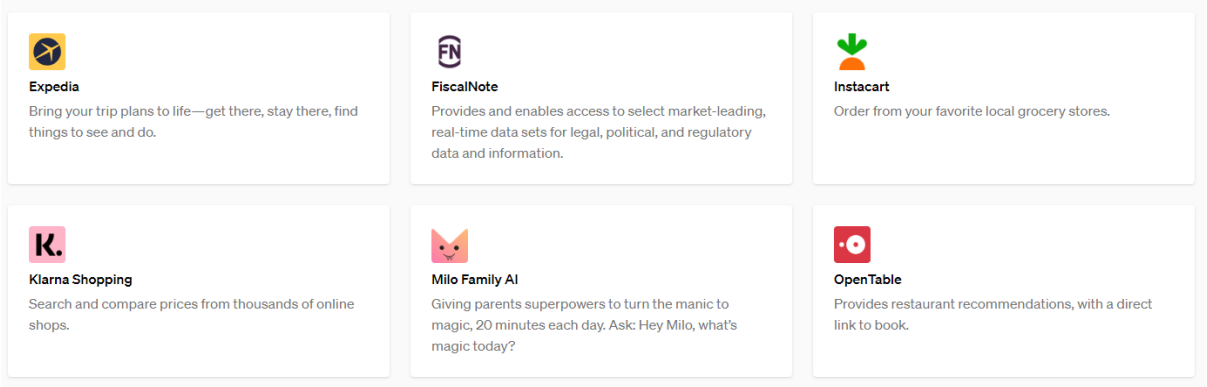
Como vemos, los plugins aparecen con la pretensión de explotar todo el potencial de este modelo cuasi humano y extenderlo a multitud de tareas, además de paliar una de sus limitaciones más importantes, el conocimiento limitado hasta el fin de su entrenamiento en 2021.

Algunos ejemplos de plugins que están en boga hoy en día y que consideramos útiles son los siguientes:

- **ChatWithPDF**: nos permite interactuar con archivos PDF como si estuviéramos hablando con un humano. Pegamos el enlace URL a cualquier archivo PDF y hacemos preguntas sobre su contenido.
- **Video Insights**: nos da la posibilidad de pegar la URL de cualquier vídeo de YouTube dentro de nuestra pregunta. Podemos pedir un breve resumen del vídeo o hacer cualquier pregunta sobre el contenido del mismo.
- **Kayak**: nos da la posibilidad de hacer preguntas sobre vuelos, hoteles, alquiler de coches y actividades con el objetivo de obtener recomendaciones personalizadas basadas en nuestras preferencias y en los datos disponibles en el conocido buscador Kayak. Puede ayudarnos a encontrar las mejores ofertas y a reservar nuestro viaje de forma rápida y cómoda.
- **KeyMate AI Search**: nos ofrece un modo de búsqueda sencillo y efectivo a través de Google. Utiliza la API de búsqueda de Google para escanear y resumir rápidamente los principales resultados de búsqueda para una palabra clave determinada. Básicamente,

analiza resúmenes de texto de todos los resultados de la primera página de búsqueda y proporciona una respuesta actualizada basada en los mismos.

- **Lexi Shopper**: nos ofrece recomendaciones de productos de Amazon. Podemos pedir a ChatGPT que nos muestre productos en función de nuestras preferencias, presupuesto o necesidades. Utiliza los datos de Amazon para acceder a millones de productos y ofrecernos enlaces y precios.
- **OpenTable**: nos facilita la búsqueda y reserva de restaurantes a nuestra petición. Podemos pedirle que nos muestre restaurantes en función de nuestra ubicación, cocina, ocasión, presupuesto u otras preferencias. A partir de la base de datos de OpenTable, que cuenta con más de 60.000 restaurantes en todo el mundo, nos ofrece recomendaciones y enlaces para reservar.
- **Zapier**: nos permite conectar y automatizar acciones dentro de varias aplicaciones web como Gmail y Slack. Por ejemplo, podemos crear una hoja de Google y añadir una nueva fila cada vez que alguien nos envíe un correo electrónico.



[The 12 Best ChatGPT Plugins]

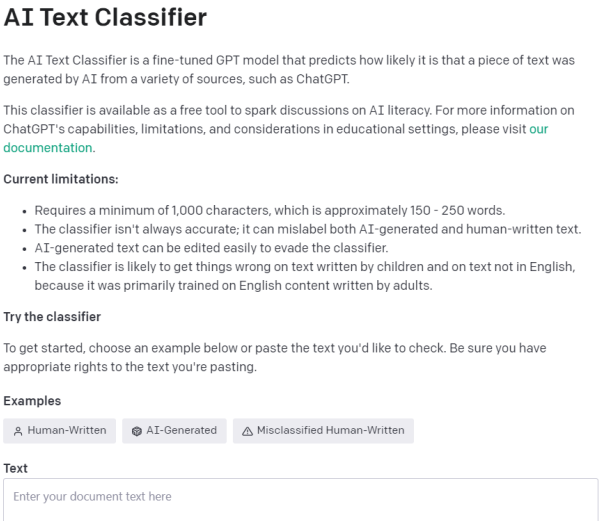
¿Podemos saber si un texto ha sido generado con ChatGPT?

Titulares similares a ***ChatGPT obliga a las universidades a replantearse el plagio*** o ***How to Ethically Use ChatGPT and Other AI Chat Bots for Assignments*** inundan internet hoy día. En multitud de ocasiones nos hacen plantearnos si, verdaderamente, podemos saber si un texto se ha generado usando ChatGPT, o cualquier inteligencia artificial generativa similar.

Gran parte de los debates relativos a la detección de texto generado por inteligencia artificial que hemos podido leer son polémicos e inconclusos. Desde hace algún tiempo, antes de la aparición de ChatGPT, ya existían herramientas que permitían realizar parafraseo y reescritura de textos con el objetivo tanto de ofrecer mejor calidad en la escritura como de mostrar un buen uso del vocabulario y de la gramática. El *copywriting* con inteligencia artificial lleva un tiempo evolucionando y mantiene una tendencia en auge con la aparición de servicios SaaS muy completos que facilitan esta labor. ¿Por qué no usar técnicas que nos permitan acelerar el proceso de escritura evitando perder tiempo en la redacción y centrándonos en la generación de valor? Este es uno de los pilares sobre los que se fundamenta este tipo de usos de las inteligencias artificiales generativas. Cuando una persona decide desarrollar sus pensamientos y plasmarlos en

un papel, puede no estar inspirado para realizar una buena redacción; sin embargo, si damos estas ideas a un modelo como ChatGPT y le pedimos que realice una reescritura en un estilo determinado, podemos conseguir paliar el efecto negativo que produciría en los lectores una mala redacción.

Si nos preguntaran con el objetivo de obtener un *sí* o un *no* como respuesta a la pregunta *¿podemos saber si un texto ha sido generado con ChatGPT?*, la respuesta corta sería, *NO*. ¿Por qué? Porque no hay ninguna herramienta que nos diga con una precisión del 100 %, con una probabilidad total de acierto, que un texto ha sido generado usando alguna herramienta de inteligencia artificial.



No obstante, OpenAI, con el objetivo de desmentir las falsas afirmaciones de que el texto generado con IA ha sido escrito por un humano, por ejemplo, en la realización de campañas automatizadas de desinformación, en el uso de herramientas de IA para plagios académicos o el posicionamiento de un chatbot de IA como si fuera un humano; ha creado una herramienta denominada ***AI Text Classifier***. Este clasificador nos indica si existe posibilidad de que un texto se haya generado usando un modelo de AI. Las clases en las que clasifica son las siguientes: ***Very unlikely to be AI-generated***, ***Unlikely to be AI-generated***, ***Unclear if it is AI written***, ***Possibly AI-generated***, ***Likely AI-generated***.

Si leemos el apartado de **limitaciones actuales** del mencionado clasificador, nos encontramos con las siguientes:

- **Requiere un mínimo de 1.000 caracteres**, que equivale aproximadamente a 150-250 palabras; esto es debido a que un texto de poca extensión puede no proporcionar suficiente contexto para que el clasificador funcione de manera óptima y produzca resultados confiables.
- **El clasificador no siempre es preciso**; puede etiquetar incorrectamente tanto texto generado por IA como texto escrito por humanos.
- El texto generado por IA puede ser editado fácilmente para **evadir al clasificador**; podemos establecer un contexto local, basado en textos escritos por nosotros, e indicar al modelo que genere texto con nuestro estilo de escritura.
- Es probable que el clasificador se **equivoque** al analizar **texto escrito por niños** y **texto que no esté en inglés**, ya que principalmente fue entrenado con contenido en inglés escrito por adultos.

[AI Text Classifier]

Tras haber realizado muchos experimentos, hemos podido observar el funcionamiento del clasificador. Por regla general, parece que los textos con estilos de escritura más pobres, sin demasiada riqueza gramatical ni semántica suelen ser clasificados como textos que es poco probable que hayan sido generados por AI. Por el contrario, si los textos que se le introducen son muy estructurados, poseen riqueza semántica y no poseen rasgos relacionados con la expresión oral suelen ser clasificados como textos que posiblemente/probablemente hayan sido generados usando AI. Podemos probar con multitud de fragmentos de textos científicos escritos años antes de que aparecieran estos modelos de generación de texto; observamos que multitud de los mismos son clasificados como textos que posiblemente hayan sido generados a través del uso de AI (falsos positivos). A pesar de lo anterior, el modelo de clasificación está sesgado con el objetivo de que produzca menos falsos positivos.

Antes de finalizar esta sección, es necesario comentar algo muy importante que no debemos obviar. Actualmente, ChatGPT puede ser entrenado en base a un contexto local, es decir, a través de la información que proporcionemos en el chat, que es nuestro espacio local para darle contexto al modelo. Esto quiere decir que podemos conseguir que adopte nuestro estilo de escritura, por ejemplo, combinando algunos textos escritos por nosotros mismos con prompts específicos. Se puede ver en el siguiente artículo un ejemplo ***How to Make ChatGPT Copy Your Writing Style***.

Conclusiones

En conclusión, la irrupción de ChatGPT en el panorama tecnológico ha demostrado que es posible mejorar múltiples aspectos de nuestra vida cotidiana. Esta innovación de OpenAI representa un hito en la historia del desarrollo de la Inteligencia Artificial, visibilizando el nivel de sofisticación al que han llegado los algoritmos de procesamiento de lenguaje natural, aprendizaje automático y las redes neuronales artificiales.

ChatGPT, con su habilidad de interactuar y comunicarse de manera coherente y contextualmente relevante, abre un abanico inmenso de posibilidades. Ya estamos presenciando su impacto en la educación, la atención al cliente, la organización y muchos otros campos, facilitando accesibilidad, eficiencia y personalización. Además, su habilidad para comprender y generar contenido en varios idiomas lo convierte en una herramienta global muy extendida.

Sin embargo, no podemos pasar por alto las implicaciones éticas y de seguridad que supone. El poder de ChatGPT para generar contenido realista también significa que existe el riesgo de su mal uso, como en la generación de desinformación o contenido dañino. Es crucial que se establezcan las políticas adecuadas para prevenir y eliminar estos riesgos. Además, hay preocupaciones sobre la privacidad y el potencial sesgo en los datos de entrenamiento.

En resumen, aunque ChatGPT trae consigo retos significativos, su potencial para impulsar avances en diversos campos es innegable. Esta tecnología disruptiva nos ofrece una nueva lente a través de la cual podemos vislumbrar el futuro de la inteligencia artificial, subrayando la importancia de un enfoque ético y considerado a medida que seguimos explorando las posibilidades de esta fascinante herramienta.

Bibliografía

- **[OpenAI (2023) Introducing ChatGPT]**
  https://openai.com/blog/chatgpt
- **[Some Glimpse AGI in ChatGPT. Others Call It a Mirage]**
  https://www.wired.com/story/chatgpt-agi-intelligence/
- **[AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference?]**
  https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks
- **[Apuntes SIGE]**
  Apuntes de la asignatura de Sistemas Inteligentes para la Gestión en la Empresa del Máster Profesional en Ingeniería Informática de la Universidad de Granada
- **[Attention Is All You Need]**
  https://dl.acm.org/doi/pdf/10.5555/3295222.3295349
- **[GPT-3]**
  https://en.wikipedia.org/wiki/GPT-3
- **[The Transformer Model]**
  https://machinelearningmastery.com/the-transformer-model/
- **[Chat GPT and GPT 3 Detailed Architecture Study-Deep NLP Horse]**
  https://medium.com/nerd-for-tech/gpt3-and-chat-gpt-detailed-architecture-study-deep-nlp-horse-db3af9de8a5d
- **[How Does ChatGPT Actually Work?]**
  https://www.scalablepath.com/data-science/chatgpt-architecture-explained
- **[How ChatGPT actually works]**
  https://www.assemblyai.com/blog/how-chatgpt-actually-works/
- **[How ChatGPT Works Technically | ChatGPT Architecture]**
  https://www.youtube.com/watch?v=bSvTVREwSNw
- **[OpenAI Models]**
  https://platform.openai.com/docs/models

- ➔ **[Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models]**
  https://arxiv.org/pdf/2304.01852.pdf
- ➔ **[ChatGPT: Everything you need to know about OpenAI's GPT-4 tool]**
  https://www.sciencefocus.com/future-technology/gpt-3/
- ➔ **[How Transformers Work]**
  https://towardsdatascience.com/transformers-141e32e69591
- ➔ **[New and improved content moderation tooling]**
  https://openai.com/blog/new-and-improved-content-moderation-tooling
- ➔ **[Pricing]**
  https://openai.com/pricing
- ➔ **[OpenAI platform]**
  https://platform.openai.com/
- ➔ **[Platform OpenAI Examples]**
  https://platform.openai.com/examples
- ➔ **[ChatGPT]**
  https://chat.openai.com/
- ➔ **[ChatGPT plugins]**
  https://openai.com/blog/chatgpt-plugins
- ➔ **[ChatGPT Prompt Engineering for Developers]**
  https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/
- ➔ **[The 12 Best ChatGPT Plugins]**
  https://approachableai.com/best-chatgpt-plugins/
- ➔ **[AI Text Classifier]**
  https://platform.openai.com/ai-text-classifier