

Tabla de Contenidos

Episodio 56: Unea Pecasto Thactet Presio. ChatGPT y los modelos de lenguaje.	1
Tabla de Contenidos	2
¿Qué es un Modelo de Lenguaje?	3
GPT	4
[G] Generative	4
[P] Pre-trained	4
[T] Transformer	4
[1-4] Historia de las versiones	5
GPT-1	5
GPT-2	5
GPT-3	6
GPT-4	6
Conceptos técnicos de GPT	7
Mecanismos de atención	7
Zero-shot Performance	7
Safeguards	8
Pendiente de resolver: el "alignment Problem"	8
Políticas empresariales y de producto	9
OpenAI - Principios	9
Políticas de uso de los modelos de Open AI	9
Privacidad	9
Entrevista de Lex Fridman a Sam Altman	9
Características de los modelos de lenguaje de OpenAI	10
1. RLHF - Reinforcement Learning with Human Feedback	10
2. Sesgos deliberados: The System Message	10
3. (consecuencia de 2) Respuestas que aclaran matices	16
4. La autocrítica: The System Card	16
Lugares muy humanos	17
Nos gustan las imperfecciones	17
Nos gusta hacer las cosas mal para explorar creativamente: pensamiento lateral	18
Somos vulnerables	18
¿Son los modelos de lenguaje (LLMs) el camino a la Inteligencia Artificial General (AGI)?	19
.....	19
Reflexiones de Jessica Flack sobre la entrevista Fridman-Altman	27
¿Los modelos de lenguaje comprenden realmente lo que les preguntamos?	28
Zachari Biondi y el sustratismo como prejuicio moral contra los robots	28
Vamos a reirnos un poco también de GPT	29
Panorama de lo que son capaces de hacer ahora los distintos tipos de IA y previsión de futuro	29
Materiales	30

Episodio 56: Unea Pecasto Thactet Presio. ChatGPT y los modelos de lenguaje.



Si, como dice el lema de este podcast, aquí reflexionamos sobre el presente, es inevitable una reflexión filosófica sobre uno de los presentes más inesperados y que más sentimientos está creando, tanto de expectación como de desazón. Nos referimos a ChatGPT, el chatbot que nos está permitiendo interactuar con algo llamado GPT, también llamado “modelo de lenguaje” y que ha puesto sobre el tablero de ajedrez filosófico una nueva bola de bollar: una nueva generación de herramientas de inteligencia artificial que nos sitúa ante viejas preguntas de una forma nueva. De eso hablaremos hoy, todo lo largo y tendido que nos permita nuestra inteligencia natural que, como ya saben, es bastante justa.

Por cierto, la imagen del episodio ha sido generada con [Stable Diffusion](#) a partir del prompt “un episodio de un podcast sobre ChatGPT”. El lema “Unea Pecasto Thactet Presio” se lo ha inventado Stable Diffusion pero nos gusta.

“La filosofía no sirve para nada” es un podcast sin pretensiones en el que reflexionaremos sobre el presente.

Participan: José Carlos García @quobit, Joaquín Herrero @joakinen@scholar.social, Juan Carlos Barajas @SociologiaDiver, Juan Antonio Torrero @jatorrero, Sergio Muñoz @smunozroncero

Fecha	29 de marzo de 2023
Participan	José Carlos García @quobit Sergio Muñoz @smunozroncero Juan Carlos Barajas @SociologiaDiver Joaquín Herrero @joakinen@scholar.social Juan Antonio Torrero @jatorrero
Descarga	Puedes descargar todos los episodios en iVoox , en Spotify , en iTunes , Google Podcasts y en nuestro canal de Telegram . Si tienes un lector de podcasts que admite enlaces RSS, este es el enlace RSS a nuestro podcast .

Sintonía	Mass Invasion , Dilo, álbum Robots (2004)
Fotos	Imagen generada por Stable Diffusion con el prompt “Un episodio de un podcast sobre ChatGPT”
Intro	Conversación con el asistente de Google
Mastodon	En @FilosofiaNada@hcommons.social publicamos noticias que nos interesan y conversamos.
Twitter	Estuvimos, pero ya casi no estamos. @FilosofiaNada
Canal Telegram	Puedes seguir la preparación de nuevos episodios suscribiéndote al canal @FilosofiaNada en Telegram
Grupo de opinión	Únete a nuestro grupo de opinión Opina FilosofiaNada para opinar sobre el episodio en preparación y enviarnos audios con preguntas o críticas con humor para nuestra intro

El efecto IA se produce cuando gente ajena al tema descarta el comportamiento de un programa de inteligencia artificial con el argumento de que no es inteligente.



Pamela McCorduck escribe: “Es parte de la historia del campo de la inteligencia artificial que cada vez que alguien descubre cómo hacer que un equipo haga algo (como jugar bien a las damas chinas o resolver problemas simples pero relativamente informales), existe un coro de críticos que dicen 'eso no es pensar'.” El investigador de IA Rodney Brooks se queja de que “cada vez que descubrimos una parte de ella, deja de ser mágica; decimos, 'Oh, eso es sólo un cálculo'.”

https://es.m.wikipedia.org/wiki/Efecto_IA

Tabla de Contenidos

- [Episodio 56: Unea Pecasto Thactet Presio. ChatGPT y los modelos de lenguaje.](#)
- [¿Qué es un Modelo de Lenguaje?](#)
- [GPT](#)
 - [\[G\] Generative](#)
 - [\[P\] Pre-trained](#)
 - [\[T\] Transformer](#)
 - [\[1-4\] Historia de las versiones](#)
 - [GPT-1](#)
 - [GPT-2](#)
 - [GPT-3](#)
 - [GPT-4](#)
 - [Conceptos técnicos de GPT](#)
 - [Mecanismos de atención](#)

- Zero-shot Performance
- Safeguards
- Pendiente de resolver: el "alignment Problem"
- Políticas empresariales y de producto
 - OpenAI - Principios
 - Políticas de uso de los modelos de Open AI
 - Privacidad
 - Entrevista de Lex Fridman a Sam Altman
 - Características de los modelos de lenguaje de OpenAI
 - 1. RLHF - Reinforcement Learning with Human Feedback
 - 2. Sesgos deliberados: The System Message
 - 3. (consecuencia de 2) Respuestas que aclaran matices
 - 4. La autocrítica: The System Card
- Lugares muy humanos
 - Nos gustan las imperfecciones
 - Nos gusta hacer las cosas mal para explorar creativamente: pensamiento lateral
 - Somos vulnerables
- ¿Son los modelos de lenguaje (LLMs) el camino a la Inteligencia Artificial General (AGI)?
- Reflexiones de Jessica Flack sobre la entrevista Fridman-Altman
- ¿Los modelos de lenguaje comprenden realmente lo que les preguntamos?
- Zachari Biondi y el sustratismo como prejuicio moral contra los robots
- Vamos a reinos un poco también de GPT
- Panorama de lo que son capaces de hacer ahora los distintos tipos de IA y previsión de futuro
- Materiales

¿Qué es un Modelo de Lenguaje?

Cita de [Kublik & Saboo, 2022](#)

El modelado del lenguaje es la tarea de asignar una probabilidad a una secuencia de palabras en un texto en un idioma específico. Los modelos de lenguaje simples pueden observar una palabra y predecir la siguiente palabra (o palabras) que probablemente la seguirán, según el análisis estadístico de las secuencias de texto existentes. Para crear un modelo de lenguaje que prediga con éxito secuencias de palabras, debe entrenarlo con grandes conjuntos de datos.

Los modelos de lenguaje son un componente clave en las aplicaciones de procesamiento de lenguaje natural. Puede pensar en ellos como máquinas de predicción estadística, donde proporciona texto como entrada y obtiene una predicción como salida. Probablemente esté familiarizado con esto por la función de autocompletar en su teléfono inteligente. Por ejemplo, si escribe "bien", la función de autocompletar podría sugerir "buenos días" o "suerte".

Antes de GPT-3, no había un modelo de lenguaje general que pudiera funcionar bien en una variedad de tareas de PNL. Los modelos de lenguaje se diseñaron para realizar una tarea específica de NLP, como la generación, el resumen o la clasificación de texto, utilizando algoritmos y arquitecturas existentes.

Un modelo de lenguaje es un tipo de inteligencia artificial que se utiliza para comprender, generar o manipular texto o habla en lenguaje natural. Estos modelos se basan en algoritmos de aprendizaje automático o aprendizaje profundo, como las redes neuronales, y se entrenan en grandes cantidades de datos de texto para aprender las estructuras y patrones del lenguaje.

Un modelo de lenguaje puede realizar varias tareas, como traducción automática, resumen de texto, generación de texto, análisis de sentimientos, respuesta a preguntas, entre otras. Los modelos de lenguaje también pueden ser específicos para un idioma o diseñados para trabajar con varios idiomas a la vez.

Un ejemplo de un modelo de lenguaje avanzado es el GPT-4 de OpenAI, que es el modelo que me impulsa. Estos modelos han demostrado un gran rendimiento en diversas tareas de procesamiento del lenguaje natural y han sido utilizados en muchas aplicaciones, desde asistentes virtuales hasta herramientas de escritura automática.

(Redactado por GPT-4)

GPT

“GPT es una herramienta, no una criatura”. Sam Altman.

[G] Generative

GPT-3, GPT-4, así como sus versiones anteriores son modelos generativos porque generan texto. El modelado generativo es una rama del modelado estadístico. Es un método para aproximar matemáticamente el mundo.

[P] Pre-trained

Un modelo previamente entrenado es entrenado (por usted o por otra persona) para una tarea más general y luego está disponible para ser ajustado para diferentes tareas. En lugar de crear un modelo desde cero para resolver su problema, utiliza el modelo entrenado en un problema más general como punto de partida y le brinda una capacitación más específica en el área de su elección utilizando un conjunto de datos especialmente seleccionado. Es posible que un modelo previamente entrenado no sea 100 % preciso, pero le evita tener que reinventar la rueda, lo que le permite ahorrar tiempo y mejorar el rendimiento.

En el aprendizaje automático, un modelo se entrena en un conjunto de datos. El tamaño y el tipo de las muestras de datos varían según la tarea que desee resolver. GPT-3 está preentrenado en un corpus de texto de cinco conjuntos de datos: Common Crawl, WebText2, Books1, Books2 y Wikipedia.

[T] Transformer

Las redes neuronales forman el núcleo del aprendizaje profundo. Su nombre y estructura están inspirados en el cerebro humano, imitando la forma en que las neuronas biológicas se envían señales entre sí. Una red neuronal es una red o circuito de neuronas que trabajan en tándem.

Las innovaciones de las redes neuronales pueden mejorar el rendimiento del modelo en tareas posteriores, por lo que los científicos de IA trabajan continuamente en nuevas arquitecturas para redes neuronales. Uno de esos inventos revolucionó la PNL tal como la conocemos hoy: el transformador. Un transformador es un modelo de aprendizaje automático que procesa una secuencia de texto de una sola vez (en lugar de una palabra a la vez) y que tiene un mecanismo poderoso para comprender la conexión entre las palabras.

Investigadores de Google y la Universidad de Toronto introdujeron la idea de un modelo de transformador en un artículo de 2017:

“Proponemos una nueva arquitectura de red simple, el Transformador, basada únicamente en mecanismos de atención, prescindiendo por completo de la recurrencia y las convoluciones. Los experimentos en dos tareas de traducción automática muestran que estos modelos son de calidad superior, al mismo tiempo que son más paralelizables y requieren mucho menos tiempo de entrenamiento.”

[1-4] Historia de las versiones

GPT-1

GPT-1 sirvió como un paso importante hacia un modelo de idioma con capacidades generales basadas en el idioma. Demostró que los modelos de lenguaje se pueden entrenar previamente de manera efectiva, lo que podría ayudarlos a generalizar bien. La arquitectura podría realizar varias tareas de NLP con muy pocos ajustes. GPT-1 usó el conjunto de datos Book Corpus, que contiene unos siete mil libros inéditos, y la parte del decodificador del transformador con autoatención para entrenar el modelo. La arquitectura se mantuvo en gran medida igual que en el transformador original. El modelo tenía 117 millones de parámetros. GPT-1 abrió caminos para modelos futuros, que podrían liberar mejor este potencial con conjuntos de datos más grandes y más parámetros. Uno de sus logros fue la capacidad de rendimiento de tiro cero decente en varias tareas de PNL.

DATASET: Book Corpus
PARAMETERS: 117 millones
ZERO-SHOT PERFORMANCE: decent

GPT-2

GPT-2 demostró que entrenar en un conjunto de datos más grande y tener más parámetros mejora la capacidad de un modelo de lenguaje para comprender tareas y superar el estado del arte de muchas tareas en configuraciones de disparo cero. También mostró que incluso los modelos de lenguaje más grandes serían aún mejores en la comprensión del lenguaje natural. Para crear un conjunto de datos extenso y de alta calidad, los autores rasparon Reddit y extrajeron datos de enlaces salientes de artículos votados en la plataforma. El conjunto de datos resultante, WebText, tenía 40 GB de datos de texto de más de ocho millones de documentos, mucho más grande que el conjunto de datos de GPT-1. GPT-2 se entrenó en el conjunto de datos de WebText y tenía 1500 millones de parámetros, 10

veces más que GPT-1. GPT-2 se evaluó en varios conjuntos de datos de tareas posteriores como comprensión de lectura, resumen, traducción y respuesta a preguntas.

DATASET: Book + WebText (40 GB de 8 millones de documentos Reddit)
PARAMETERS: 1.500 millones de parámetros
ZERO-SHOTS: en VARIAS tareas no entrenadas

GPT-3

En la búsqueda por construir un modelo de lenguaje aún más sólido y poderoso, OpenAI creó el modelo GPT-3. Tanto su conjunto de datos como el modelo son aproximadamente dos órdenes de magnitud más grandes que los utilizados para GPT-2: GPT-3 tiene 175 000 millones de parámetros y se entrenó en una combinación de cinco corpus de texto diferentes, un conjunto de datos mucho más grande que el que se usó para entrenar GPT- 2. La arquitectura de GPT-3 es en gran medida la misma que la de GPT-2. Se desempeña bien en tareas de NLP posteriores en configuraciones de disparo cero y pocos disparos. GPT-3 tiene capacidades como escribir artículos que son indistinguibles de los artículos escritos por humanos.

Mientras que GPT-2 mostró algunas capacidades de disparo cero en tareas posteriores, **GPT-3 puede llevar a cabo tareas aún más novedosas cuando se le presenta un contexto de ejemplo.**

Los investigadores de OpenAI encontraron notable que simplemente escalar los parámetros del modelo y el tamaño del conjunto de datos de entrenamiento condujo a avances tan extraordinarios. En general, son optimistas de que estas tendencias continuarán incluso para modelos mucho más grandes que GPT-3, lo que permitirá siempre- modelos de aprendizaje más potentes capaces de aprender con pocos o ningún intento simplemente ajustando con precisión un tamaño de muestra pequeño. Mientras lee este libro, los expertos estiman que probablemente se están construyendo e implementando modelos de lenguaje basados en un billón de parámetros. Hemos entrado en la era dorada de los grandes modelos de lenguaje, y ahora es el momento para ti.

DATASET: 5 corpora
PARAMETERS: 175.000 millones
ZERO-SHOTS: OK
FEW-SHOTS: OK

GPT-4

[GPT-4 Web Page](#)
[GPT-4 Technical Report](#)
[GPT-4 System Card](#)

Este es el video en el que se transmitió en directo la primera demo de GPT-4.

Conceptos técnicos de GPT

Mecanismos de atención

Los mecanismos de atención son una parte fundamental de los modelos Transformer, que son un tipo de arquitectura de redes neuronales artificiales utilizadas para tareas de procesamiento de lenguaje natural (NLP, por sus siglas en inglés).

En un modelo Transformer, la atención se utiliza para determinar la importancia relativa de diferentes partes de la entrada (por ejemplo, una oración) para generar una representación contextualizada de la misma. En particular, la atención permite a la red enfocarse en partes relevantes de la entrada para cada tarea específica.

En los modelos Transformer, existen tres tipos de mecanismos de atención:

- **Atención de entrada (self-attention):** Este tipo de atención permite que la red se centre en diferentes partes de la entrada en función de su relevancia para la tarea. Por ejemplo, para un modelo de traducción de idiomas, el modelo puede centrarse en diferentes partes de la oración de entrada en función de la posición de la palabra objetivo en la oración de salida.
- **Atención de salida (encoder-decoder attention):** Este tipo de atención se utiliza en modelos de traducción y otros modelos secuenciales, y permite que el modelo se centre en diferentes partes de la entrada y la salida en función de su relevancia para la tarea.
- **Atención de posición (positional attention):** Este tipo de atención se utiliza para incorporar información de posición en la representación contextualizada de la entrada.

(Redactado por GPT-3.5)

Zero-shot Performance

El zero-shot performance, o rendimiento sin entrenamiento previo, se refiere a la capacidad de un modelo de lenguaje natural para realizar una tarea específica sin haber sido entrenado explícitamente en esa tarea.

Por ejemplo, si se entrena un modelo de lenguaje para la tarea de clasificación de sentimientos en inglés, y luego se le da una oración en español para clasificar, el modelo podría ser capaz de hacerlo utilizando su comprensión general del lenguaje y las características universales de los sentimientos. Esto se consideraría un ejemplo de rendimiento sin entrenamiento previo, ya que el modelo no ha sido entrenado específicamente en la tarea de clasificación de sentimientos en español.

Los modelos que logran un buen rendimiento sin entrenamiento previo se consideran muy útiles, ya que pueden ser más adaptables a nuevos dominios y lenguas que los modelos que sólo han sido entrenados en una tarea específica. Además, el zero-shot performance es una medida importante para evaluar la capacidad de generalización de los modelos de lenguaje natural, lo que es esencial para que estos modelos puedan ser utilizados en una variedad de aplicaciones prácticas.

El one-shot performance y el some-shot performance son medidas similares al zero-shot performance, pero difieren en el grado de entrenamiento previo que se permite en el modelo.

El one-shot performance se refiere a la capacidad de un modelo para realizar una tarea específica

después de haber sido entrenado en esa tarea sólo una vez, es decir, con un solo ejemplo. Por ejemplo, si se entrena un modelo de lenguaje para la tarea de identificación de entidades en textos médicos, y se le da un solo ejemplo de una oración que contiene una entidad médica, el modelo podría ser capaz de identificar esa entidad en otras oraciones similares sin necesidad de más entrenamiento.

El some-shot performance se refiere a la capacidad de un modelo para realizar una tarea específica después de haber sido entrenado en esa tarea con un número limitado de ejemplos. Por ejemplo, si se entrena un modelo de lenguaje para la tarea de traducción de inglés a español, y se le da sólo un pequeño conjunto de oraciones en inglés y sus correspondientes traducciones en español, el modelo podría ser capaz de traducir oraciones en inglés que no están en su conjunto de entrenamiento con cierto grado de precisión.

En resumen, mientras que el zero-shot performance mide la capacidad de un modelo para realizar una tarea sin entrenamiento previo, el one-shot performance y el some-shot performance miden la capacidad de un modelo para realizar una tarea con un grado limitado de entrenamiento previo.

(Redactado por GPT-3.5)

Safeguards

Las safeguards son medidas de seguridad diseñadas para proteger los sistemas de inteligencia artificial (IA) y las aplicaciones basadas en IA de posibles riesgos y amenazas, como el sesgo, la discriminación, la privacidad y la seguridad. Estas medidas se implementan para garantizar que los sistemas de IA sean justos, éticos, seguros y confiables.

Las safeguards pueden incluir diversas técnicas, como la validación cruzada para reducir el sesgo en los datos de entrenamiento, la evaluación sistemática de los modelos para detectar y corregir errores, la limitación del acceso a los datos de entrenamiento y la implementación de medidas de privacidad para proteger los datos de los usuarios.

Además, las safeguards también pueden involucrar la transparencia y la explicabilidad de los modelos de IA, lo que significa que los desarrolladores deben proporcionar información clara y accesible sobre cómo se toman las decisiones en el modelo y cómo se procesan los datos.

Las safeguards son importantes para garantizar que la IA se utilice de manera responsable y ética, y para fomentar la confianza y la adopción de la tecnología. La implementación de medidas de seguridad adecuadas puede ayudar a mitigar los riesgos asociados con la IA y garantizar que se utilice para el beneficio de la sociedad en general.

(Redactado por GPT-3.5)

Pendiente de resolver: el "alignment Problem"

El "alignment problem" (problema de alineación) en el contexto de ChatGPT y otros modelos de lenguaje de inteligencia artificial se refiere al desafío de garantizar que los objetivos, valores e intereses del sistema de IA estén alineados con los de los seres humanos. En otras palabras, se busca que la IA entienda, interprete y responda a las necesidades y expectativas de los usuarios de manera efectiva, ética y segura.

Los modelos de lenguaje como ChatGPT son el resultado del entrenamiento en grandes conjuntos de datos de texto que contienen información de diversas fuentes y perspectivas. Como resultado, pueden adoptar sesgos, creencias erróneas y comportamientos no deseados que no siempre están en consonancia con las intenciones del usuario o el bienestar general.

El problema de alineación en IA plantea varios desafíos:

- **Sesgo y justicia:** Asegurar que el modelo no tenga sesgos sistemáticos o discriminación en sus respuestas.
- **Seguridad:** Prevenir que el modelo proporcione respuestas dañinas, ofensivas o inapropiadas.
- **Privacidad:** Evitar que el modelo divulgue información confidencial o sensible.
- **Robustez y fiabilidad:** Garantizar que el modelo responda de manera consistente y confiable a las entradas del usuario, incluso en casos de entradas ambiguas o maliciosas.
- **Control de contenido y políticas:** Establecer pautas claras y efectivas para el comportamiento del modelo y garantizar que se adhiera a ellas.

Resolver el problema de alineación es crucial para el desarrollo y adopción segura y ética de sistemas de inteligencia artificial como ChatGPT en diferentes aplicaciones y contextos.

(Redactado por GPT-4)

Políticas empresariales y de producto

OpenAI - Principios

<https://openai.com/charter>

Políticas de uso de los modelos de Open AI

<https://openai.com/policies/usage-policies>

Privacidad

[Addressing criticism, OpenAI will no longer use customer data to train its models by default](#)

[Employees Are Feeding Sensitive Business Data to ChatGPT Technical Report](#)](<https://arxiv.org/abs/2303.08774>)

Entrevista de Lex Fridman a Sam Altman

Características de los modelos de lenguaje de OpenAI

1. RLHF - Reinforcement Learning with Human Feedback

El Reinforcement Learning with Human Feedback (RLHF) es una técnica de aprendizaje por refuerzo que involucra la interacción entre un agente de aprendizaje automático y un humano experto. A diferencia del aprendizaje por refuerzo tradicional, donde el agente aprende de la retroalimentación proporcionada por el entorno, en RLHF el agente recibe retroalimentación directamente del humano experto.

En RLHF, el agente aprende a través de la iteración de dos fases: en la primera fase, el agente lleva a cabo una tarea y recibe retroalimentación del experto humano, que puede ser positiva o negativa. En la segunda fase, el agente utiliza la retroalimentación recibida para ajustar su comportamiento y mejorar su desempeño en la tarea.

RLHF se ha utilizado en una variedad de aplicaciones, como en la enseñanza asistida por computadora, donde un agente de aprendizaje interactúa con un profesor humano para mejorar su desempeño en una tarea educativa, o en la robótica, donde un robot puede recibir retroalimentación de un humano experto para mejorar su capacidad para realizar una tarea específica.

El RLHF es una técnica prometedora en el aprendizaje por refuerzo, ya que permite una interacción más natural y eficiente entre humanos y agentes de aprendizaje automático, lo que puede mejorar la calidad del aprendizaje y hacer que las aplicaciones de la IA sean más útiles y accesibles para una variedad de usuarios y contextos.

(Redactado por GPT-3.5)

Caben las siguientes preguntas:

- ¿Qué criterio sigue el humano que corrige a la IA durante su entrenamiento?
- ¿Cómo se representa durante la fase RLHF las diferencias culturales y de criterio entre personas?

Sam Altman comenta en la entrevista con Fridman que este es aún un tema en investigación y uno de los puntos preocupantes porque ¿qué criterios hay que representar?

2. Sesgos deliberados: The System Message

El "system message" es un mecanismo deliberado de introducción de sesgos por el cual GPT puede adoptar cierto punto de vista en particular, sesgado si se quiere, para adaptar culturalmente su mensaje. Es el usuario el que tiene el volante y elige qué versión sesgada de la respuesta quiere.

Por este mismo mecanismo GPT puede adoptar un rol en su respuesta: responde como Chiquito de la Calzada

El contexto de una conversación es un conjunto de hechos que se quieren que GPT tenga en cuenta a la hora de elaborar una respuesta.

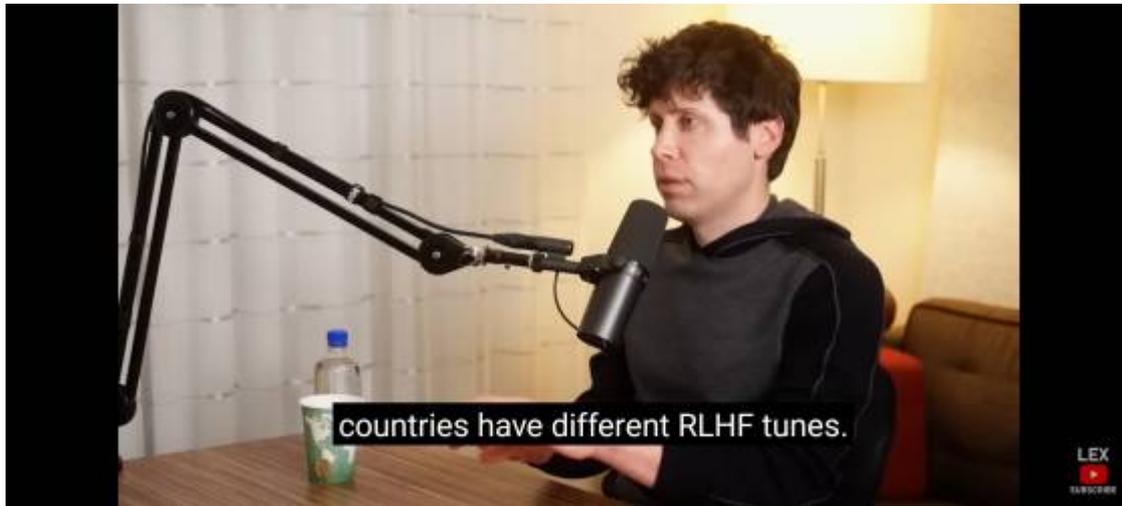
También se usa para poder contextualizar una nueva pregunta de la misma conversación ya que

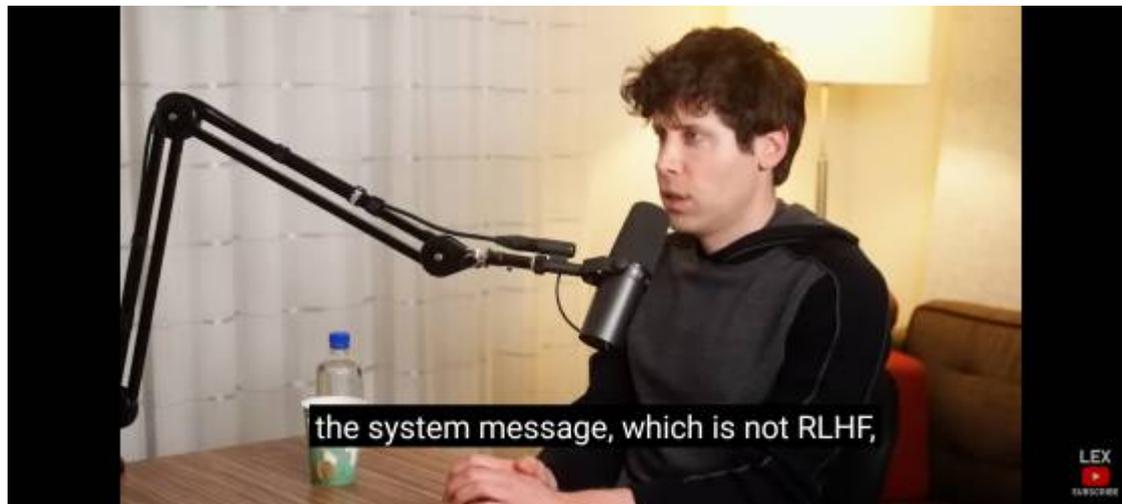
todas las preguntas y respuestas de esa conversación pasan a ser parte del contexto.













3. (consecuencia de 2) Respuestas que aclaran matices

Sam Altman: ChatGPT ha devuelto los matices a la conversación en internet. Ha devuelto aquello que Twitter nos quitó.

4. La autocrítica: The System Card

El documento llamado "The System Card" es una herramienta utilizada en el campo de la inteligencia artificial (IA) para describir y documentar un sistema de IA específico. Este documento también se conoce como "tarjeta de identificación del sistema" o "ficha técnica del sistema" en español.

El término "system card" se deriva del hecho de que el documento tiene una estructura similar a una tarjeta de identificación, con información detallada y resumida sobre el sistema de IA en cuestión.

La idea detrás de "The System Card" es proporcionar una descripción detallada del sistema de IA que permita a los desarrolladores, usuarios y reguladores comprender mejor cómo funciona el sistema, cómo se ha entrenado y cómo se desempeña en diferentes situaciones. La tarjeta puede incluir información como el tipo de algoritmos de aprendizaje automático utilizados, el conjunto de datos utilizado para entrenar el modelo, la precisión y el rendimiento del modelo, y cualquier otro detalle relevante sobre el sistema de IA.

El objetivo de la "The System Card" es aumentar la transparencia y la comprensión de los sistemas de IA, especialmente en el contexto de aplicaciones críticas, como la atención médica, la seguridad y la justicia. Al proporcionar una descripción detallada y estandarizada del sistema de IA, se espera que la tarjeta ayude a garantizar que los sistemas de IA sean justos, éticos y confiables para su uso en una variedad de aplicaciones.

(Redactado por GPT-3.5)

[GPT-4 System Card by OpenAI - March 15, 2023](#)

Esta tarjeta de sistema analiza GPT-4, el último modelo de lenguaje grande en la familia de modelos GPT. Desde que terminó la capacitación en agosto de 2022, hemos estado evaluando, probando adversariamente y mejorando iterativamente el modelo y las mitigaciones a nivel del sistema a su alrededor. Nuestras mitigaciones y procesos alteran el comportamiento de GPT-4 y evitan ciertos

tipos de usos indebidos, aunque tienen limitaciones, lo que apunta a la necesidad de una planificación y gobernanza anticipadas y una mayor investigación de seguridad. Nuestro enfoque de implementación equilibra la minimización del riesgo de la implementación, la habilitación de casos de uso positivos y el aprendizaje de la implementación. Los modelos GPT a menudo se entrenan en dos etapas.

- Primero, se les entrena, utilizando un gran conjunto de datos de texto de Internet, para predecir la siguiente palabra.
- Luego, los modelos se ajustan con datos adicionales, utilizando un algoritmo llamado aprendizaje de refuerzo a partir de la retroalimentación humana (RLHF), para producir resultados que son preferidos por los etiquetadores humanos.

El entrenamiento de modelos de lenguaje en grandes conjuntos de datos de texto ha dado lugar a capacidades como el **aprendizaje de pocos intentos** (some-shot performance) y la capacidad de llevar a cabo una amplia gama de tareas de lenguaje natural que abarcan diferentes dominios, incluida la respuesta a preguntas, la aritmética y la clasificación. El ajuste fino ha hecho que estos modelos sean más controlables y útiles.

Alucinaciones

Cital del “GPT-4 Technical report”:

Despite its capabilities, GPT-4 has similar limitations to earlier GPT models [1, 37, 38]: it is not fully reliable (e.g. can suffer from “hallucinations”), has a limited context window, and does not learn from experience. Care should be taken when using the outputs of GPT-4, particularly in contexts where reliability is important.

Lugares muy humanos

Nos preocupa que el aumento en las competencias de las IAs relegue al humano de lugares que, de alguna manera, pensamos que son “naturalmente nuestros”.

Por eso no está de más pensar en algunos lugares que, por el momento, siguen siendo reductos humanos. Para bien o para mal.

Nos gustan las imperfecciones



Nos gusta hacer las cosas mal para explorar creativamente: pensamiento lateral

Somos vulnerables

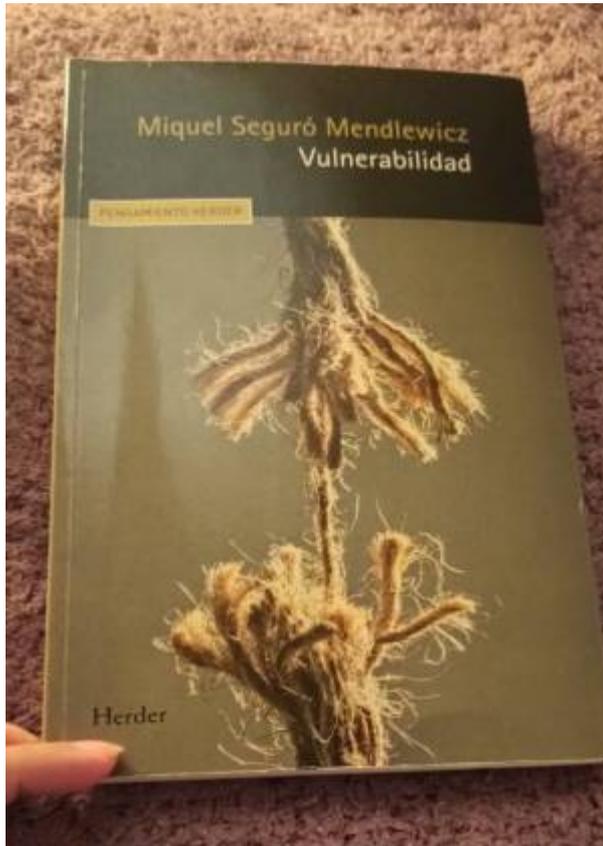
Ha sido un placer escuchar en el [@ateneodemadrid](#) el debate sobre "Vulnerabilidades" entre [@miquel_seguro](#), [@ClaudineLefebvre](#) y Javier Moscoso, con Mercedes López como maestra de ceremonias. Ser vulnerable o afectable es parte de la condición humana y merece la pena ser pensado. pic.twitter.com/2cZLbATeyv

— Ibis Albizu (@IbisAlbizu) [March 28, 2023](#)

Recomiendo esta lectura de [@miquel_seguro](#) que presenta a un Descartes único relacionándolo con un tema de total actualidad. La permanente duda sobre la verdad que acompaña a la filosofía

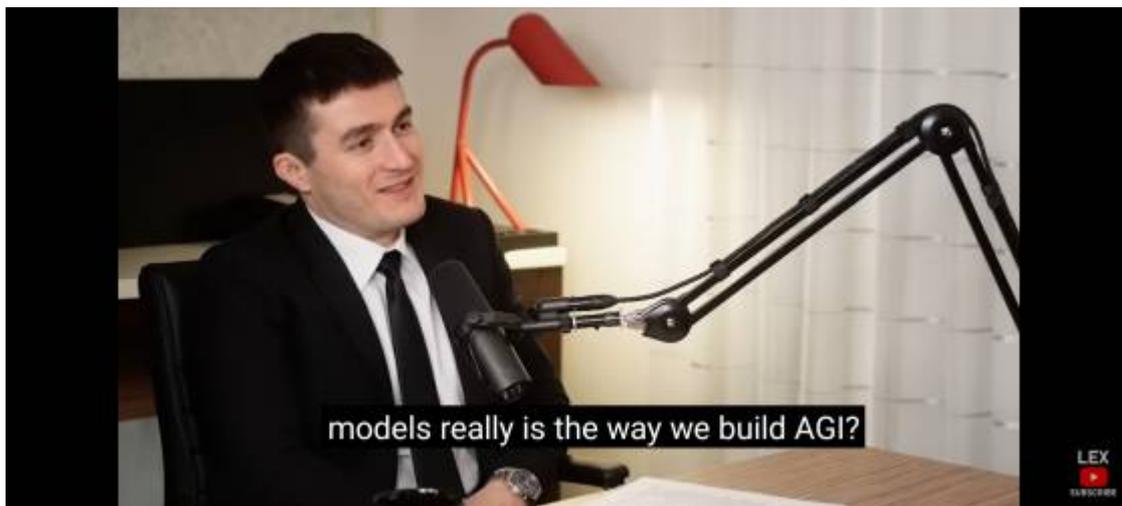
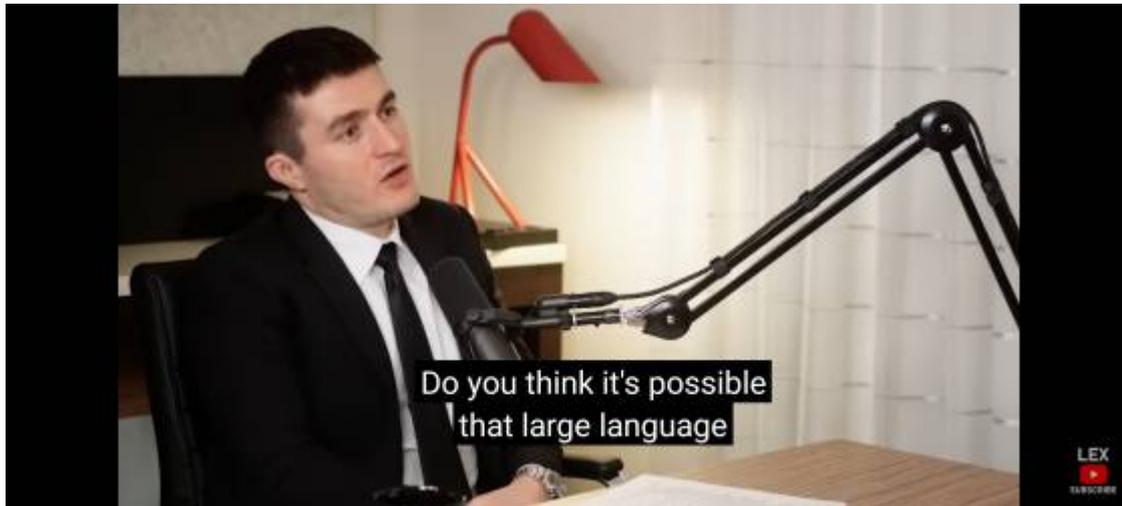
es una saludable forma de mostrarse vulnerable. Publicado en [@herderbarcelona](https://twitter.com/DonZeAYyBS).
pic.twitter.com/DONZeAYyBS

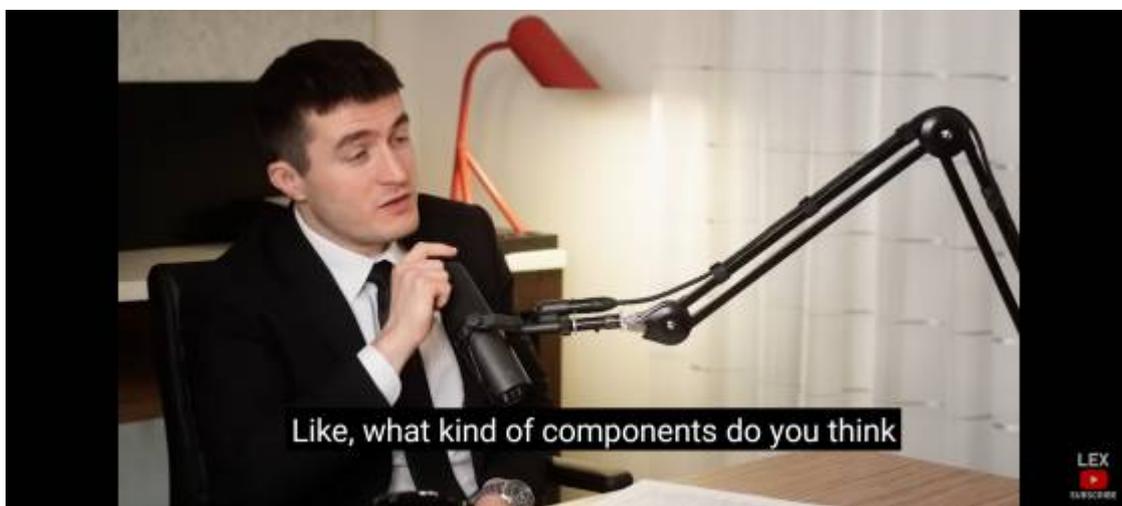
— Ibis Albizu (@IbisAlbizu) [March 28, 2023](#)



¿Son los modelos de lenguaje (LLMs) el camino a la Inteligencia Artificial General (AGI)?

Lex Fridman plantea si no es un requisito para alcanzar la Inteligencia Artificial General la capacidad de experimentar el mundo sensorialmente y, por tanto, ser capaces tanto de sentir dolor (tal como propone Antonio Damasio) como de intervenir en el mundo para comprenderlo (como propone Judea Pearl). De estos temas ya estuvimos hablando en el [Episodio 17: Es posible que la IA sea sensible. No hay quien vea si será una buena idea #Porvivir](#)

















Reflexiones de Jessica Flack sobre la entrevista Fridman-Altman

En [este hilo de Twitter](#), Jessica Flack, científica de datos, bióloga evolutiva y profesora en el [Santa Fe Institute](#) hace las siguientes reflexiones:

(Traducción automática del hilo, revisada)

La discusión ofrece una perspectiva útil y equilibrada. Preguntas/comentarios:

1. ¿Es una medida de AGI el grado en que es robusto para Artisanal Prompt Engineering (APE), es decir, que puede inferir lo que el usuario quiere saber independientemente de lo bueno que sea el prompt?
2. ¿Qué podría hacer posible teorizar sobre el conocimiento deseado por el usuario? ¿Qué pasa si (opcionalmente) se incluye con cada solicitud un paquete de datos que resume al usuario? Por supuesto, esto no necesita ser un atajo con un resumen, pero los resúmenes permitirían a los usuarios controlar cómo se presentan.
3. La inteligencia LLM parece ser puramente estadística en lugar de mecanicista... integradores del conocimiento humano sin entendimiento. Dado esto, la cuestión de la alineación de la IA no tiene sentido, al menos hasta que construyamos una IA mecanicista. Hay que invertir en capacitar a humanos para obtener mejores LLM.
4. ¿Es el debate sobre el 'resbalón' de los LLM y que produzcan un resultado sesgado realmente un debate sobre cómo resumimos lo que piensan los humanos y qué hay en ese resumen que pueda constituir 'conocimiento'?
5. Un tema importante para muchas disciplinas es la diferencia entre mecanicista + estadística. Dicho de otra manera, ¿cuándo un modelo estadístico comienza a sentirse mecanicista? ¿Es suficiente tener, por ejemplo, un posterior + un anterior, que implican, entre otras cosas, escalas de tiempo?
6. Los modelos de lenguaje grande no son más que dispositivos de auto reconocimiento de espejo para humanos, que reflejan nuestro rastro digital colectivo que, al servicio de capturar la esencia a través de la compresión con pérdida, a veces parece distorsionar la llamada verdad.
7. Esperar la verdad de un modelo diseñado para capturar la esencia es malinterpretar cómo se calcula la salida a partir de la entrada (el mapeo) y es un error. Los LLM (hasta ahora) + AI generalmente no tienen autonomía o agencia, en realidad no alucinan y tienen inteligencia

dudosa.

8. A pesar de todo parece probable que los humanos que usan LLM cambien drásticamente la organización económica y la naturaleza del trabajo.

¿Los modelos de lenguaje comprenden realmente lo que les preguntamos?

Preguntamos a GPT-4:

¿Tú comprendes las cosas que te pregunto?

Sí, como inteligencia artificial, mi función principal es comprender y procesar preguntas para proporcionar respuestas y asistencia en una amplia variedad de temas. Si tienes alguna duda o inquietud, no dudes en preguntarme y estaré encantado de ayudarte.

¿Pero entiendes de la misma forma que los humanos?

Aunque puedo procesar y comprender el lenguaje humano de manera efectiva, mi comprensión es diferente a la de un ser humano. Como inteligencia artificial, no experimento emociones, pensamientos subjetivos ni consciencia de la misma manera que las personas. Mi capacidad para comprender se basa en el análisis de patrones y la correlación de datos en lugar de la experiencia humana vivida. Esto significa que, aunque puedo proporcionar información útil y relevante en respuesta a tus preguntas, no experimento el proceso de comprensión de la misma manera que un ser humano lo haría.

Sobre este asunto es muy interesante el artículo **“Does GPT-4 Really Understand What We’re Saying? One question for David Krakauer, an evolutionary theorist and president of the Santa Fe Institute for complexity science.”** disponible en esta dirección:

<https://nautil.us/does-gpt-4-really-understand-what-were-saying-291034/>

También es muy relevante el artículo **“ChatGPT Is a Blurry JPEG of the Web. OpenAI’s chatbot offers paraphrases, whereas Google offers quotes. Which do we prefer?”**, del escritor estadounidense Ted Chiang, disponible en

<https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web>.

Zachari Biondi y el sustratismo como prejuicio moral contra los robots

A partir de [este hilo en Twitter](#) en el que citaba la cuestión de la “máquina moral” comentada por el filósofo [Zachary Biondi](#) nos ha interesado su crítica acerca de que la reflexión filosófica sobre las máquinas y la moralidad está lastrada por un prejuicio que favorece a los seres hechos de carbono (y no todos), al que él llama “substratism” y que relaciona con otras formas de opresión.

¿En qué punto una pieza de tecnología es miembro de la comunidad moral? ¿Cuándo empieza a tener

derechos un robot?

Los materiales en los que profundizar son varios, algunos conocidos, y, en todo caso, cortos de leer/escuchar:

- [Stop Talking about the Moral Status of Machines](#), artículo de Zach Biondi en The VIM Blog ([traducción automática](#))
- [The Coming Oppression](#), artículo de Zach Biondi en The VIM Blog ([traducción automática](#))
- [VIMcast #51](#), When Will Machines Deserve Rights?, episodio de podcast en el que Zach Biondi comenta los dos artículos anteriores

En [la entrevista que Lex Fridman hace a Stephen Wolfram](#) se habla de la “inteligencia extraterrestre” como un concepto que ya estamos consiguiendo superando así el prejuicio hacia las formas biológicas de inteligencia: la IA que estamos construyendo en máquinas.

Vamos a reirnos un poco también de GPT

En los últimos meses han ido apareciendo IAs cada vez más potentes. Sus creadores intentan mantener a raya los usos menos éticos o respetuosos, pero no siempre lo consiguen... El último modelo en sufrirlo –de momento– ha sido ChatGPT: un chatbot con habilidades aparentemente prodigiosas y unas salvaguardas de contenido inapropiado que se pueden desactivar con un poco de imaginación, y conseguir que te diga cómo robar un coche o una receta para una tarta de caca. En esta charla explicaremos las últimas novedades en este campo y pondremos ejemplos de cómo sacar partido a estos modelos, pero también de los sesgos que poseen, los problemas éticos que plantean y las tontunas que podemos hacerles decir.

Charla presentada en [T3chfest 2023](#)

Panorama de lo que son capaces de hacer ahora los distintos tipos de IA y previsión de futuro

Hace un año hablar de IAs que podían generar caras realistas de personas que no existen, o de chatbots super elocuentes capaces de redactar párrafos de texto sin equivocarse, era para muchos motivo de asombro. Pero esto ha cambiado. Recientemente, con la llegada de herramientas como ChatGPT, DALL-E 2, Codex, Stable Diffusion y otras tantas, la Inteligencia Artificial se ha vuelto “mainstream”, y una herramienta más en nuestro día a día.

Todo el mundo habla ya de Inteligencia Artificial y muchos empiezan a entender el potencial de una tecnología llamada a cambiar el mundo. Siendo así quiero que te vuelvas a dejar sorprender y que, más allá del hype y de las expectativas exageradas, juntos respondamos la pregunta: ¿qué podemos esperar del futuro de la IA, cuando parece que el futuro ya es hoy?

Charla presentada en [T3chfest 2023](#)

Materialiales

https://www.youtube.com/watch?v=L_Guz73e6fw

<https://www.youtube.com/watch?v=PoDNIEEw9bU>

<https://youtu.be/eURA2pB-rAU>

<https://youtu.be/YqjyYZ4jsAw>

GPT-3: Building Innovative NLP Products Using Large Language Models, Kublik & Saboo, 2022

https://books.google.es/books/about/Gpt_3.html?id=QFi0zgEACAAJ&redir_esc=y

Sparks of Artificial General Intelligence: Early experiments with GPT-4

From:

<http://filosofias.es/wiki/> - **filosofias.es**

Permanent link:

<http://filosofias.es/wiki/doku.php/podcast/episodios/56>

Last update: **2023/03/30 17:43**

