ChatGPT y el fin de un mundo

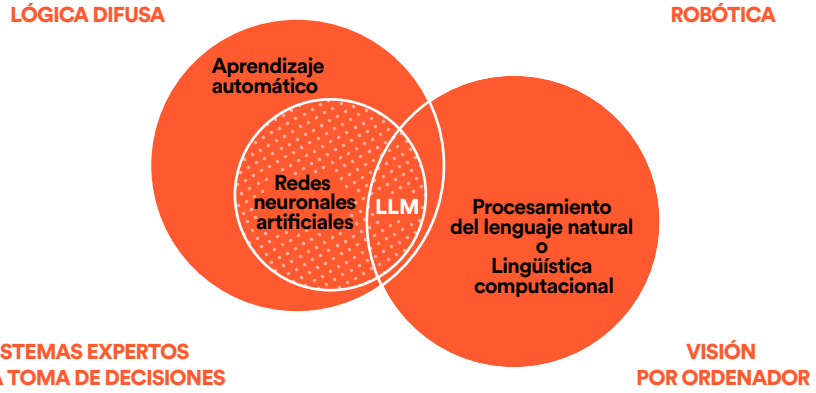
EPISODIO 1

¿QUÉ SON LOS LLM?

Raquel Fernández
Catedrática de Lingüística Computacional
Institute for Logic, Language and Computation
Universidad de Ámsterdam

Herramientas como ChatGPT se han hecho muy populares últimamente. ChatGPT es un ejemplo de «modelo de lenguaje de gran tamaño», o *large language model* (LLM) en inglés. Este tipo de tecnología existe desde hace unos años dentro del campo de la lingüística computacional, que es una subdisciplina de la inteligencia artificial.

Inteligencia artificial



Los modelos de lenguaje de gran tamaño más recientes son capaces de generar automáticamente textos que se parecen mucho a los textos producidos por seres humanos.

¿CÓMO LO HACEN?

Un LLM es

Un modelo de lenguaje es un programa de ordenador diseñado por humanos (investigadores en IA, en lingüística computacional, etc.).

Está basado en un modelo matemático denominado ***red neuronal artificial***.

Se trata de un programa que utiliza la técnica de **aprendizaje automático** (*machine learning*): un LLM aprende a generar textos y se entrena, para generarlos, por medio de datos (¡muchos datos!).

El tipo de red neuronal artificial que utilizan los LLM actuales se denomina *transformador*.

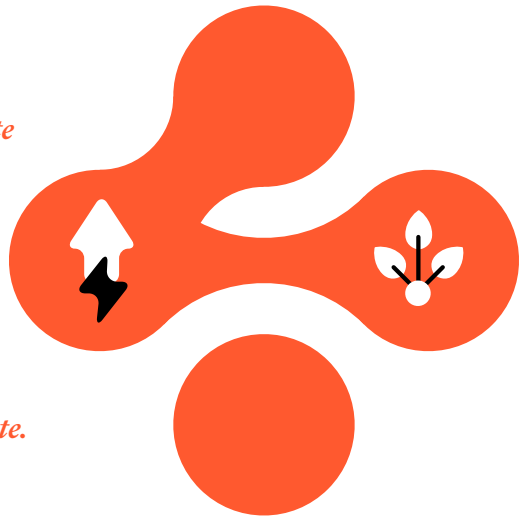
LA SIGLA ***GPT*** HACE REFERENCIA A LA EXPRESIÓN INGLESA *GENERATIVE PRE-TRAINED TRANSFORMER*

¿Por qué se llaman *modelos de lenguaje «de gran tamaño»*?

Se llaman «de gran tamaño» porque las redes neuronales artificiales que hay detrás de los LLM tienen una estructura compleja, con miles de millones de **parámetros**, G que son como las palancas que controlan la respuesta del sistema.

Además, necesitan grandes cantidades de material de aprendizaje para poder aprender a producir textos de buena calidad.

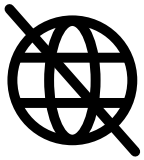
Para entrenar un modelo de este tipo, hacen falta miles de ordenadores trabajando en paralelo durante meses. Hay que tener en cuenta que esto requiere un consumo energético considerable, que puede tener consecuencias negativas para el medioambiente.



¿Cuál es el *material de aprendizaje*?

El **material de aprendizaje** (*training data*) es un conjunto de textos que tiene que ser tan grande y diverso como sea posible. Lo más habitual es utilizar textos descargados de internet, por ejemplo de Wikipedia, de libros y periódicos digitales, de blogs o de redes sociales, como Twitter, Instagram o Facebook.

En principio, un modelo de lenguaje no está directamente conectado a internet: durante la fase de entrenamiento, solo tiene acceso a los textos descargados que forman el material de aprendizaje.



Se debe tener en cuenta que el uso de este tipo de datos puede ser problemático: por una parte, puede haber problemas relacionados con privacidad, consentimiento y propiedad intelectual; por otra parte, puede ser que los datos estén sesgados y no representen correctamente ciertos sectores de población.

¿Qué aprende a hacer un LLM?

La idea fundamental es muy sencilla: un LLM aprende a **predecir la palabra siguiente** en una frase o texto, dado un contexto inicial o *prompt*. Ⓖ

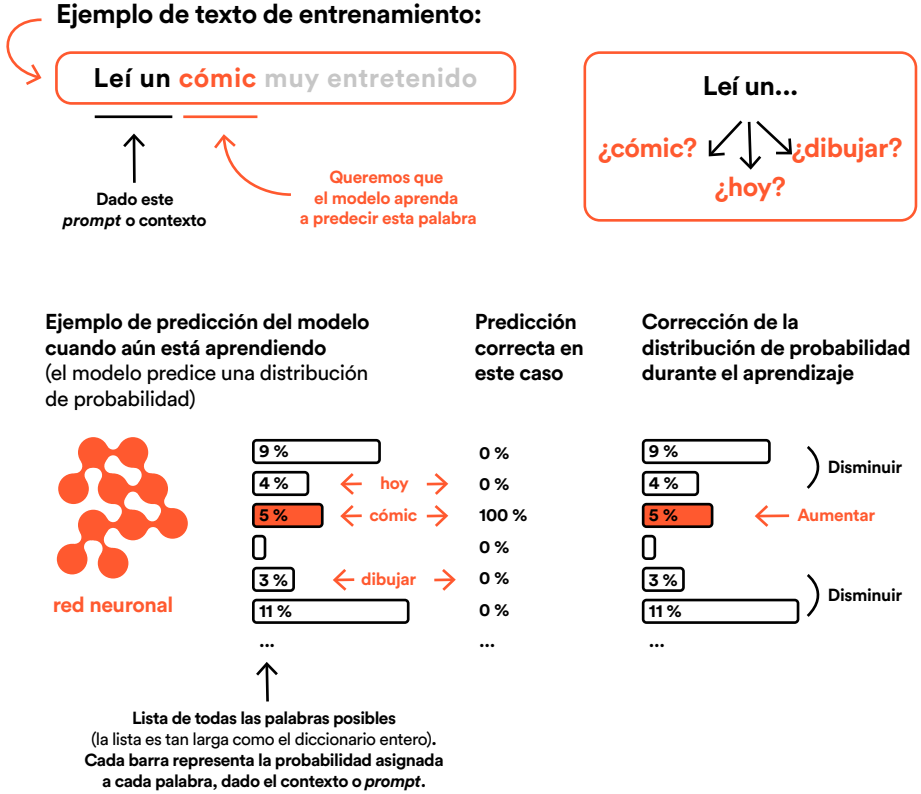
Por ejemplo, no todas las palabras del diccionario tienen la misma posibilidad de aparecer después de un fragmento de texto como «Hoy hace...»: las palabras *sol, exactamente, quince* o *mucho* son posibles continuaciones, mientras que palabras como *casa*, *bailar* o *es* tienen una probabilidad muy baja (casi nula) de ser la palabra siguiente.

A partir de todas las frases y textos que forman parte del **material de aprendizaje**, el modelo aprende a asignar las probabilidades adecuadas a cada palabra posible y, por extensión, a cada frase.

Es decir, aprende los patrones que determinan qué frases son naturales y razonables en un contexto dado o como respuesta a una consulta.

Predecir una palabra tras otra permite generar textos, respuestas y conversaciones enteras. Por ello, los LLM se consideran un ejemplo de **inteligencia artificial generativa**.

Aprendizaje no supervisado



¿Cómo se entrena un LLM con ChatGPT?

ChatGPT ha sido creado por una empresa privada, OpenAI, que no ha revelado todos los detalles sobre cómo ha entrenado este modelo concreto. Aun así, sabemos que el entrenamiento incluye tres fases:

En una primera fase, todos los LLM se entrenan para predecir la palabra siguiente dado un contexto de entrada.

Esto se considera **aprendizaje no supervisado**, porque el modelo aprende a partir de grandes cantidades de textos existentes, sin intervención humana adicional.

Los LLM como ChatGPT pasan por dos fases de aprendizaje más.

En una segunda fase, trabajadores humanos producen ejemplos de conversaciones entre dos personas que simulan conversaciones entre una persona y ChatGPT.

Estas conversaciones se utilizan como material de aprendizaje adicional. El objetivo sigue siendo aprender a predecir la palabra siguiente, pero ahora en un contexto de diálogo interactivo.

Cabe mencionar que las condiciones de trabajo y las remuneraciones de estos trabajadores no son óptimas.

Finalmente, en una tercera fase, se utiliza el aprendizaje supervisado: cuando ya es posible interactuar con el modelo, las respuestas generadas automáticamente por ChatGPT son evaluadas por más trabajadores, que las etiquetan como «buenas» o «malas» (por ejemplo, las respuestas con lenguaje ofensivo o sobre temas controvertidos se etiquetarán como «malas»).

Esta información se utiliza como *feedback* con el objetivo de potenciar respuestas que se adecuen a las preferencias humanas. Este tipo de aprendizaje se denomina ***aprendizaje de refuerzo a partir de las preferencias humanas***, o *reinforcement learning with human feedback* (RLHF) en inglés.

¿Podemos confiar en los textos y respuestas generados por un LLM?

Hay que tener presente que los modelos de lenguaje **no incorporan ningún mecanismo que asegure la veracidad del contenido** que producen: las frases que generan pueden sonar totalmente plausibles y razonables, pero no hay garantía de que sean ciertas (a veces lo serán, y a veces, no).

El hecho de que los textos generados sean gramaticales y tengan una forma natural puede dar lugar a la **antropomorfización** de los *chatbots* basados en modelos de lenguaje.

Es decir, a menudo asignamos cualidades humanas (como inteligencia, razonamiento lógico o emociones) a máquinas que solo aparentan tenerlas. Sin embargo, estos rasgos no forman parte de las capacidades de un modelo de lenguaje.

Preguntas de reflexión

1. ¿Qué aspecto de todo lo aprendido aquí sobre los LLM te ha llamado más la atención?

2. ChatGPT es un ejemplo de LLM. ¿Conoces más? Bing AI, Bard... ¿Te suenan? ¿Qué pueden hacer? La tecnología avanza rápidamente. Mientras trabajamos en la publicación de este capítulo, ChatGPT ha anunciado su próxima conexión a internet (en principio, un LLM no está conectado a internet) y Bing ha anunciado una extensión para la generación de imágenes. ¿En qué punto estamos en el momento en que lees esto?

3. ¿Qué papel juegan la lógica y la coherencia en la evaluación de las respuestas generadas por ChatGPT? ¿Qué harías para distinguir entre una respuesta válida y una incorrecta?

4. Algunas voces alertan sobre el peligro de asignar características humanas a una inteligencia artificial. ¿A ti te parece peligroso? ¿Por qué?

5. La intervención de las personas en la creación, entrenamiento y respuesta de los LLM es importantísima. ¿Puedes dar algunos ejemplos? ¿Consideras que la implicación de las personas/agentes humanos es suficientemente agradecida, valorada, apreciada? ¿Por qué crees que es así? ¿Qué implicaciones tiene que sea así?

6. Si utilizas LLM para tus deberes, ¿consideras una obligación mencionarlo? ¿Por qué?

G GLOSARIO

- **ANTROPOMORFIZACIÓN**: Otorgar, a un animal o cosa, características o motivaciones humanas.
- **PARÁMETROS**: Valores numéricos que definen el comportamiento del modelo.
- ***PROMPT***: Instrucción, pregunta o frase inicial proporcionada al modelo de lenguaje para guiar su respuesta o generación de texto.
- **RED NEURONAL ARTIFICIAL**: Modelo inspirado en el funcionamiento del cerebro humano. Está formado por un conjunto de nodos, conocidos como *neuronas artificiales*, que están conectados y transmiten señales los unos a los otros. Estas señales se transmiten desde la entrada hasta generar una salida.

REFERENCIAS PARA SABER MÁS

a. Bender, Emily M. et al. «On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?» en *Proceedings of FAccT 2021*, p. 610-623 (2021, dl.acm.org)
b. Brown, Tom B. et al. «Language Models are Few-Shot Learners» en *Advances in Neural Information Processing Systems* (p. 1877-1901). Curran Associates (2020, neurips.cc)
c. Pointon, Chris. «The carbon footprint of ChatGPT» en *Blogpost sobre el impacto medioambiental de ChatGPT* (2022, medium.com)
d. «Circuito OEIAC sobre el uso responsable y sostenible del ChatGPT» (2023, udg.edu)
e. «Beijing Consensus on Artificial Intelligence and Education» (2019, unesco.org)
f. Sabzalieva, Emma y Valentini, Arianna. «ChatGPT e Inteligencia Artificial en la educación superior. Guía de inicio rápido» (2023, unesco.org)

Texto: Raquel Fernández / Diseño: La Puput Gràfica Coop V
Fundación "la Caixa", 2023