

Los modelos de lenguaje no escriben el texto de una sola vez sino fragmento a fragmento. A cada paso necesitan volver a leerlo todo: la consulta, el contexto original y lo escrito hasta ese momento. A medida que el texto crece, al modelo le cuesta cada vez más encontrar las instrucciones iniciales, y, en un efecto *bola de nieve*, los fragmentos que va generando pueden incluir errores o imprecisiones. Esta es una de las causas de que, a veces, los modelos fallen y proporcionen datos falsos. Son las *alucinaciones* de la IA.

[FOTOGRAFÍA: TATIANA IKOEVA / [ISTOCK](#)]

Es posible que alguna vez al ir a su médico haya encontrado a un sustituto que debe ponerse al día y ojear su historial. Imaginemos que ese sustituto fuera una persona distinta cada vez que le hace una nueva pregunta. Dejando a un lado que la propuesta parece sacada de una pesadilla o de un sketch de humor absurdo, para poder contestarnos con eficacia, esa persona no solo tendría que mirar el historial médico, sino también una transcripción de la conversación que vamos manteniendo. Esta extraña situación describe con bastante precisión lo que sucede cuando hablamos con un *chatbot* de inteligencia artificial.

En una conversación con un *chatbot* el modelo de lenguaje que lee la consulta y genera la respuesta no cambia su estado en el proceso, es decir, no modifica su configuración interna. Tampoco incorpora la consulta a su memoria. De hecho, el modelo de lenguaje no tiene memoria y su estado es inmutable.

Entonces, ¿cómo nos responde sin perder el hilo?

La clave está en que a la consulta que le hacemos se le añade el historial de la conversación. Esto permite que cada consulta pueda ser respondida por distintas copias del modelo de lenguaje o por la misma. No obstante, aunque nos respondiera siempre la misma copia, nos recordaría todavía menos que Dory, la despistada pez de la película Buscando a Nemo.

El contexto se queda pequeño

Al conjunto de elementos que se le proporcionan al modelo de lenguaje junto con la consulta se le denomina [contexto](#) e incluye los mensajes anteriores de la conversación. Aunque también puede incorporar documentos adjuntos, resultados de búsquedas de información, etc.

El contexto es una herramienta clave para que los chatbots funcionen adecuadamente. Sin embargo, tiene un tamaño limitado y, cuando se llena, el sistema debe decidir qué conservar y qué descartar. Para ello debe resumir la conversación sin perder la información relevante. También se puede dotar a los modelos de lenguaje de un mayor contexto, pero esta vía tiene inconvenientes:

1. Un contexto más grande, a la larga, también acaba llenándose.
2. Usar un contexto mayor requiere una mayor capacidad de cómputo, lo que encarece el sistema y empeora los tiempos de respuesta.
3. Los modelos de lenguaje son menos eficaces a la hora de encontrar información relevante cuando el contexto es muy grande. Este último problema también nos sucede a las personas, pero resulta curioso que nos sucede además de forma muy similar.

Sesgos para recordar

Las personas tenemos sesgos cognitivos para recordar mejor los primeros elementos de una lista o las primeras impresiones, pero también lo más reciente. Se les conoce respectivamente como sesgo de primacía y sesgo de [recencia](#) (la tendencia a recordar mejor la información presentada al principio o al final, respectivamente, de una secuencia o conversación). Estos sesgos condicionan cómo recordamos y deben ser tenidos en cuenta para establecer una comunicación eficaz.

El estudio de Nelson F. Liu, de la Universidad de Stanford, y otros autores muestra que los modelos de lenguaje también tienen sesgos similares. Su mecanismo de atención, que así se llama, capta mejor la información en el inicio y el final del contexto, y se degrada en la parte central del mismo. ¿Por qué sucede así?

La información situada al final del contexto se tiene en cuenta porque es la última que ha procesado el modelo justo antes de empezar a generar la respuesta. Sin embargo, de forma contraintuitiva, estos modelos de lenguaje también aprovechan mejor la información que aparece al comienzo del contexto.

Una posibilidad que sugiere el estudio es que el entrenamiento del modelo le ha hecho aprender a fijar más atención al comienzo de los textos, donde suelen aparecer instrucciones o información clave que sirven para interpretar lo que viene después.

El problema de generar textos largos

Otra cuestión sobre el contexto que está recibiendo mucha atención recientemente es el de los problemas a la hora de escribir respuestas extensas.

Es posible que haya observado que a los modelos de lenguaje les cuesta mantener las instrucciones si tienen que generar un texto muy largo. Esto es crítico si se les ha solicitado escribir un informe o el código fuente de un programa informático, donde la coherencia debe mantenerse a lo largo de todo el texto generado.

Un grupo de investigadores de la Singapore University of Technology and Design, encabezados por Yuhao Wu, han propuesto un banco de pruebas que consiste en varias tareas de generación de textos largos que mantienen una coherencia temporal (un diario personal y los menús de un restaurante, ambos a lo largo de un año) o espacial (diseñar un rascacielos y la planificación urbanística de una barriada).

La tarea encomendada debe seguir además unas instrucciones precisas de distinto tipo: introducir determinada información en un punto preciso (por ejemplo, el rascacielos debe tener un gimnasio en la planta 34), meter información en un segmento o rango determinado (por ejemplo, las plantas de la 1 a la 9 deben alojar un centro comercial), o repetir determinada información a intervalos predeterminados (por ejemplo, a partir de la planta 20, cada diez plantas debe haber un jardín).

El estudio muestra que ninguno de los modelos evaluados hizo las tareas de forma satisfactoria hasta el final y que todos empeoraban su rendimiento cuanto más largo era el texto que se les pedía. Sin embargo, con la suficiente concentración, las personas sí somos capaces de tener en cuenta este tipo de instrucciones. ¿Por qué los modelos de lenguaje no?

¿Cómo pierde el hilo la IA?

Estos modelos no suelen estar entrenados para mantener las instrucciones en textos muy largos, lo cual puede explicar que su rendimiento se degrade en tareas que requieren generar textos extensos.

La razón es que los modelos de lenguaje no escriben el texto completo, de una sola vez, sino fragmento a fragmento (esos fragmentos no son exactamente palabras, aunque eso no es relevante aquí). Para generar cada fragmento tienen en cuenta la consulta, el contexto original y lo escrito hasta ese momento. Es decir, a cada paso, necesitan volver a leerlo todo.

Sin embargo, a medida que el texto crece hay más información compitiendo por su atención y es ahí donde surgen los problemas: al modelo le cuesta cada vez más encontrar las instrucciones que se le dieron. Además, los fragmentos que va generando pueden incluir errores o imprecisiones que, cuando el modelo vuelve a leerlos, se amplifican al generarse otros nuevos fragmentos, causando un efecto *bola de nieve*. Esta, de hecho, es una de las causas de que, a veces, los modelos fallen estrepitosamente, proporcionando datos falsos. Es lo que popularmente se conoce como [alucinaciones de la IA](#).

Estos problemas nos muestran que, pese a su capacidad y velocidad para procesar y generar textos, y pese al avance que están experimentando, los modelos de lenguaje pueden equivocarse en tareas sencillas [por ignorar aspectos clave de nuestras peticiones](#). Así que conviene conocer sus limitaciones y anticiparnos a ellas para poder sacar el mejor rendimiento posible a los modelos de lenguaje.