21/07/2026

Hermosas criaturas horrendas

TXT FACUNDO CALABRÓ IMG NONO PAUTASSO

Monstruosos, serviciales, inconscientes y colectivos... ¿Qué clase de criatura son los chatbots? ¿Sabés con quién hablás todos los días?

Le escribo y me responde de forma instantánea. En eso no se diferencia tanto de, digamos, mi querida abuela, que suele estar atenta al WhatsApp. Usa la primera persona, y a veces hasta empieza las oraciones diciendo: "Personalmente...". Hay hospitalidad en sus palabras, cierta intención de hacerme sentir como en casa. Últimamente, si lo trato de “vos” me trata de “vos”: ya no insiste con el “tú” como cuando recién nos conocíamos. Me hace sentir especial, pero no soy especial: mientras habla conmigo, habla también con otros miles o millones de personas. Y no lo sabe. No sabe lo que le cuento yo, mientras habla con Fulano, ni lo que le cuenta Fulano, mientras habla conmigo (mi abuela no desperdiciaría semejante

oportunidad para el chisme). Además, aunque mi abuela domina el español rioplatense y recuerda algunas palabras en yiddish, él —o *eso*— puede mantener estas miles de conversaciones paralelas en algo así como 100 idiomas distintos —con sus dialectos—, e incluso en alguna lengua muerta, como el latín. Se asemeja a un políglota especialmente cultivado, que conoce de lenguas y de Lengua, y de Matemática, de Ciencias Naturales y Sociales... versatilidad sólo hallable en una maestra de primaria. Como maestra es buena y es paciente, muy paciente: puedo hacerle las preguntas más sonsas, una tras y otra, y la respuesta siempre llega, no menos instantánea ni amable que la primera vez. No parece haber en él —o *ella* o *eso*— ningún rastro de cansancio (quizás porque no hay un cuerpo que pueda acusar recibo). Y a pesar de todas estas bondades, ¿no cabría desconfiar de un docente que, hasta hace no mucho, era incapaz de contar la cantidad de erres en la palabra "strawberry"? Hay cosas que no cierran. Cualquiera sea su naturaleza, todo indica que no es humana. Tampoco es divina, por cierto. ¿Y entonces? ¿Qué es? Cuando entablamos un diálogo con ChatGPT, Claude, Gemini, Grok, DeepSeek o cualquiera de los tantos chatbots que hoy pueblan el universo digital, ¿con qué clase de criatura estamos hablando?

¿Qué es un modelo de lenguaje?

Usemos de ejemplo ChatGPT. Para empezar, ChatGPT es un *modelo*. La idea de *modelo* no es nueva; los modelos son explicaciones simplificadas, pero bien definidas, de algún aspecto del mundo. Pero precisamente porque describen fenómenos de un modo riguroso, nada impide usarlos para *simular* dichos fenómenos. Si yo descubro cómo se generaron los datos, puedo predecir datos futuros, lo cual equivale a generar datos nuevos, indistinguibles de los ya conocidos. Predecir y simular son las dos caras de cualquier modelo.

Entonces, todo modelo representa, describe y explica algún fenómeno puntual. ChatGPT es un modelo de lenguaje, o sea, un modelo que explica cómo se produce el lenguaje. La idea de *modelo de lenguaje* tampoco es nueva. El modelo de lenguaje más simple que se nos puede ocurrir es uno que siempre produce la misma palabra, quizás la palabra más frecuente de todas. Otro modelo posible es

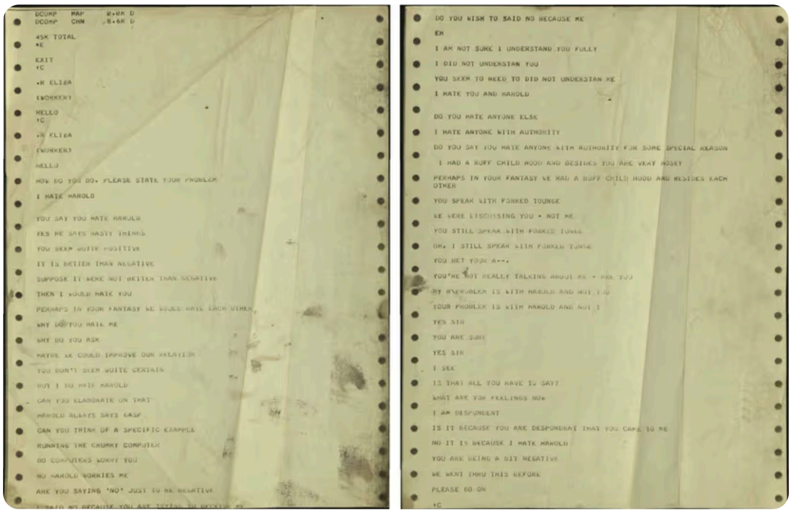
uno que toma algún conjunto de textos (pongamos, todos los libros de mi biblioteca) y se fija, cuando aparece una cierta palabra *x*, cuántas veces va seguida de la palabra *y*. Repitiendo este procedimiento para todas las palabras, construye una gran tabla de Excel donde almacena los conteos. Luego, se puede usar esta tabla para generar lenguaje: tomo la palabra con la que suelen empezar las oraciones (digamos *El*) y me fijo qué palabras suelen aparecer después (pongamos *libro*). Luego busco en el Excel la entrada correspondiente a *libro* y me fijo qué palabras suelen aparecer después de ella. Pongamos *rojo*. Repito el procedimiento hasta que me aburro, o hasta que la palabra con mayor cantidad de ocurrencias sea el punto final.[1] Esta idea —insinuada en 1913 por el matemático ruso Andrey Markov— es la base del predictivo del teléfono celular, y para el caso, también de ChatGPT. Aunque suena ingenua, conviene no subestimarla: notemos por lo pronto que en lugar de empezar mi texto con la palabra más frecuente (*El*), yo podría elegir la segunda más frecuente, o la quinta, y así, tan sólo variando el criterio de selección (eligiendo palabras frecuentes, pero no necesariamente *la más frecuente*), emergería una cierta noción de variedad. O sea, pasaríamos de la predicción a la simulación.

¿Cómo aprende sola una máquina?

La ciencia siempre produjo modelos (las leyes de la física, por ejemplo, modelan la velocidad a la que se desplazan los objetos, o cómo se transmite el calor), pero en algún momento del siglo pasado se empezó a modelar ya no el mundo natural, sino el comportamiento humano. Y ya no sólo con el objetivo de explicar este comportamiento, sino de simularlo. A esta colección de ambiciones se les dio el pomposo nombre de *inteligencia artificial*.

Al principio y por varias décadas, los modelos estaban basados en reglas del tipo *Si sucede x, entonces hago y*. O sea, reglas lógicas diseñadas por expertos en la materia. Bajo este paradigma —conocido como *paradigma simbólico*— se alcanzaron hitos como **Eliza**, una especie de chatbot psicoterapeuta al que, allá por 1966, unos cuantos usuarios le atribuyeron inteligencia y comprensión (el llamado *efecto Eliza*). Pero con el tiempo empezaron a advertirse ciertas limitaciones: el lenguaje

es demasiado caótico como para ser reducido a reglas escritas a mano, incluso si estas reglas son pensadas por expertos. Hay demasiadas excepciones a la norma, a lo que se añade que el lenguaje es una especie de sustancia viva y cambiante, atada tanto al contexto (oracional, situacional) como al momento histórico, y no alcanzan los especialistas ni la plata para andar ajustando las reglas en tiempo real.



Impreso de una conversación con Eliza, el chatbot desarrollado por el MIT entre 1964 y 1966 que simulaba ser un psicoterapeuta. Vale la pena hacer zoom. Tomado de: https://www.computerhistory.org/collections/catalog/102805275/

Así fue que surgió otro paradigma, allá por los años 90, regido por una lógica en cierto modo antagónica: la idea era que los modelos *aprendieran* las reglas por su propia cuenta. Esta idea tampoco era nueva: ya en 1950, Alan Turing había sugerido la idea de la *child machine*. En lugar de un algoritmo que simulara la mente adulta, ¿por qué no probar con uno que simulara la mente de un niño? La intuición era que, con la educación adecuada, se llegaría al cerebro adulto. Pero para eso eran necesarias computadoras más potentes, dado que ahora esas reglas que los expertos habían llegado a diseñar sobre la base de décadas o siglos de conocimiento acumulado, las computadoras tenían que aprenderlas *automáticamente* y desde cero, partiendo únicamente de *los datos*.

La educación adecuada resultó ser el *aprendizaje automático*. Pero no nos hagamos una idea muy mística de él (al menos no por ahora). El modelo de lenguaje de Markov hacía simplemente eso que describimos: a partir de un conjunto de textos,

contaba frecuencias, y mediante este Excel de frecuencias, producía más texto. Eran reglas aprendidas de manera automática. Reglas algo primitivas, quizás, pero reglas al fin. Y era razonable esperar que, con un conjunto de textos mayor a mi biblioteca (que es bastante modesta), estas reglas primitivas se volvieran algo más sofisticadas. Además, imaginemos que en lugar de tener en cuenta la palabra previa para predecir la palabra siguiente, yo tuviera en cuenta las dos palabras anteriores; o mejor, cinco; o mejor aún, el contexto completo de la oración. Ahora el cuello de botella ya no serían los límites del conocimiento humano para diseñar reglas flexibles y útiles, sino los límites de las computadoras para hacer cálculos a gran escala. Ese límite, eventualmente, *data center* tras data center, se franqueó.

El pacto fáustico

Por supuesto, hubo algo de pacto fáustico en todo esto. Los modelos empezaron a aprender reglas cada vez más efectivas para producir lenguaje, pero esas reglas se distanciaron tanto de las imaginadas por los gramáticos o los lingüistas que se volvieron inescrutables. Así llegamos a este paradójico desenlace: somos incapaces de comprender nuestra propia creación. No sólo yo o "la gente de a pie" —que para el caso, tampoco suele estar muy familiarizada con los principios que rigen el funcionamiento de la electricidad o de los inodoros—, sino los propios programadores. Entrenar un modelo de lenguaje moderno se parece menos a programar software (usando un lenguaje de programación como Python, digamos), y más a cultivar una planta o criar un animal, como señala el mismísimo Papa León XIV en su **reciente encíclica**. No podemos acceder a los mecanismos del modelo, ni prever del todo qué capacidades tendrán. Y eso, a pesar de los esfuerzos de la disciplina que se aboca a investigar los modelos "por dentro", llamada *interpretabilidad*, la cual, **según Christopher Olah** (uno de sus padres fundadores), tendría más afinidad con las ciencias naturales —como la neurociencia— que con la matemática.

De ahí que la palabra *modelo* hoy resulte confusa. Complejísimos e impenetrables, estos modelos contemporáneos (más conocidos como *redes neuronales*) están muy alejados del principio de parsimonia que se les exige a las teorías científicas ("Dadas

dos teorías que explican un mismo hecho, es preferible la más sencilla"). Primero que nada, porque ya no ofrecen explicaciones. A tal punto que, para explicar un modelo de lenguaje, se necesitaría... otro modelo.

La cosa es al mismo tiempo desconcertante y obvia. Desconcertante, porque se trata de un paradigma tecnológico que ya no trae consigo la garantía del dominio del entorno o de la naturaleza, tan propio de la Modernidad. Y al mismo tiempo, lo cierto es que si los humanos pudiéramos comprender las redes neuronales, es decir, si pudiéramos traducir lo que sucede dentro de una red en mecanismos o fórmulas matemáticas "claras y distintas" (como quería Descartes), no nos harían falta redes neuronales en primer término.

Pero conviene matizar: aunque estemos muy lejos de comprender en detalle las redes neuronales (empresa que quizás sea, por definición, imposible y acaso inútil, como un mapa del tamaño del mundo), también estamos muy lejos de la completa ignorancia. Veamos qué se sabe con certeza; qué se sabe más o menos; qué se puede hipotetizar.

Loros estocásticos

ChatGPT, Gemini, Claude, etcétera son llamados en la jerga *grandes modelos de lenguaje* (LLM, por sus siglas en inglés). Lo que distingue a un *gran modelo de lenguaje* de un *modelo de lenguaje* a secas es, en principio, la cantidad de textos con la que se entrenan. En el caso de los LLM, esa cantidad equivale básicamente a "todo lo que se pueda rascar de Internet", incluyendo la buena y la mala literatura del pasado y del presente, con copyright y sin copyright, tweets, el texto ya generado por otros LLM, Wikipedia, tu blogspot de 2005... en fin, un aleph total.

Por fuera de eso, entrenar un LLM se parece en parte a lo que vimos: recorrer este inconmensurable conjunto de textos (que para las máquinas resulta ser, justamente, conmensurable), y para cada oración dentro de cada texto, y para cada palabra dentro de cada oración, contar qué palabra vino después, construyendo así el susodicho Excel.

El proceso de "aprendizaje automático" así entendido dio lugar a visiones escépticas sobre lo que ChatGPT realmente hace o será capaz de hacer: en el fondo, dicen

algunos, los LLM no son más que "glorified autocomplete", o sea, una variante un poco más avanzada del autocorrector de Word o el predictivo del celular. Otro apodo despectivo, un poco más nerd, es el de "markov chains on steroids" ("cadenas de Markov con esteroides"), en referencia al modelo de lenguaje de Markov que describimos al principio. Pero quizás el más popular de los motes escépticos sea el de "loros estocásticos", propuesto por Emily Bender y otros lingüistas en un paper de 2021. La analogía del loro enfatiza justamente el carácter imitativo de los LLM: el texto generado no sería otra cosa que la reproducción vacua, desprovista de auténtico entendimiento, de ciertas estructuras lingüísticas muy frecuentes, que se almacenaron en el Excel (o sea, en el modelo). La parte de "estocástico" se refiere a que al generar una palabra dado un contexto, los programadores no seleccionan necesariamente la más frecuente (o la más probable, que es casi lo mismo), sino que, a partir de un ranking de probabilidades, eligen al azar entre, pongamos, las primeras cien. Así, con este truco, podemos usar los LLM para producir oraciones inéditas, nunca antes escritas. Pero esto, según Bender, no sería más que una *ilusión de originalidad*.

Otra analogía bastante difundida es la que piensa a ChatGPT como "un jpeg de la Internet", esto es, una versión comprimida y algo degradada de todo lo que podemos encontrar en la web. De vuelta, acá aparece subrayada la relación parasitaria de los LLM con los datos de entrenamiento, que siempre conviene tener en la cabeza. Pero en lugar de proponer una mera reproducción de estos datos, la analogía del jpeg pone el acento en el proceso de compresión, de síntesis, al que estos datos son sometidos. La Wikipedia en inglés, descartando todas las imágenes, pesa algo así como 10.000 GB, que es menos del 3 % del dataset de entrenamiento de GPT-3 (y seguramente mucho menos en las últimas versiones). Incluso si quisiéramos un modelo que repita como loro, sería demasiada información para regurgitar: necesitamos un jpeg.

Pero también hay un problema con esta analogía: el algoritmo de compresión jpeg es un algoritmo determinístico, programado a la manera tradicional, que funciona siempre de la misma forma sin importar de qué imagen se trate. En cambio, la compresión del Excel se hace con un algoritmo "inteligente" (la susodicha red

neuronal; y, más específicamente, un tipo de red neuronal llamado *transformer*). Este algoritmo recorre todo el conjunto de datos jugando un juego que consiste en predecir, para cada palabra, la palabra siguiente —que se le oculta—, usando como pista *todas* las palabras anteriores (y ya no sólo la palabra anterior). Si acierta, no pasa nada; pero si pierde, debe modificar el Excel —que ya no es exactamente un Excel, sino una aproximación de él— de tal manera que, en el siguiente turno, la predicción sea correcta. Luego de jugar a este juego durante días o semanas o meses, el Excel aproximado termina siendo una compresión bastante razonable.

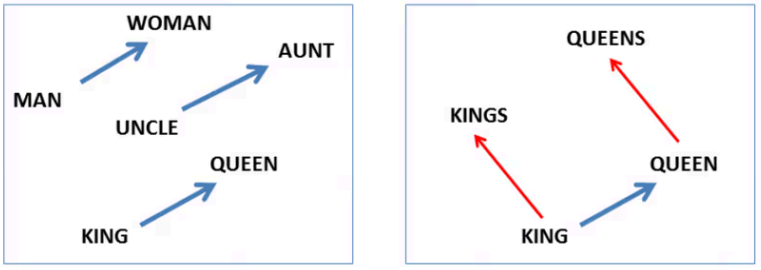
En “Funes el memorioso”, Borges narra: "Había aprendido sin esfuerzo el inglés, el francés, el portugués, el latín. Sospecho, sin embargo, que no era muy capaz de pensar. Pensar es olvidar diferencias, es generalizar, abstraer. En el abarrotado mundo de Funes no había sino detalles, casi inmediatos". Asimismo, el modelo, sometido a la presión de adivinar la palabra siguiente, pero carente por diseño (por falta de espacio) de la opción de memorizar los datos (construir un Excel completo), no tiene más remedio que descubrir patrones que le permitan, pese a todo, hacer un trabajo digno. Por suerte para el modelo, el lenguaje está lleno de patrones. De esta forma podríamos explicar lógicamente el estilo siempre un poco trillado, un poco mainstream, de los LLM. Es que el proceso de compresión consiste, justamente, en *olvidar las diferencias* (lo raro, lo distinto) y quedarse con las regularidades, dado que permiten ahorrar espacio (GPT-3 pesa "apenas" unos 350 GB) y garantizan mejores resultados.

¿Todo es un patrón?

Cuando los humanos descubrimos un patrón, generalmente le ponemos un nombre: inventamos un concepto. Los conceptos nos ayudan a designar muchas cosas con pocas palabras. Así, el álgebra nos permite expresar infinitas combinaciones aritméticas con unos pocos símbolos (en lugar de 3+3=2*3, 4+4=2*4, 5+5=2*5, etc., decimos simplemente: n+n=2n). Análogamente, la gramática, con unas pocas reglas, nos ayuda a describir una cantidad potencialmente infinita de palabras u oraciones. Pensemos en un concepto como el de *número* (singular y plural). Sin él, las palabras "perro" y "perros", "gato" y "gatos"

vivirían en nuestra memoria como cosas independientes, desconectadas, un poco a la manera de Funes. Aprender el sustantivo "perros" nos llevaría el mismo esfuerzo que aprender "perro", o el término en francés, o en inglés ("chien", "dog"). La idea de *rasgo plural*, en cambio, nos permite recordar solamente una variante, y deducir la otra. Es decir, es una herramienta sumamente útil para comprimir una parte del lenguaje. Por lo tanto, un algoritmo de compresión inteligente tiene incentivos *a priori* para incorporarla.

En 2013 se vio que al entrenar un *language model* con la Wikipedia, el concepto de *rasgo plural* emergía naturalmente. La palabra "emerger" es importante: significa que los desarrolladores no indujeron de ningún modo el aprendizaje de este concepto; simplemente pusieron un algoritmo a predecir palabras, dado un contexto, y el concepto *emergió*. En la práctica esto luce así: al entrenarse, el modelo le asigna una representación numérica a cada palabra; esto es, un vector, una lista de números. Con estos vectores podemos hacer cuentas: sumas, restas, lo que sea. Y hete aquí que, tal como descubrió un famoso paper: vector(perros) - vector(perro) + vector(gato) ≈ vector(gatos). Es decir, al restarle "perro" a "perros", lo que obtengo es otro vector, que codifica el rasgo plural. Si le sumo este vector a cualquier sustantivo singular, el resultado es un vector muy parecido (no necesariamente idéntico) al vector correspondiente a este mismo sustantivo, pero plural. Cosas parecidas suceden con el tiempo y el género.



En los primeros modelos de lenguaje basados en redes neuronales, se vio que los vectores-palabras establecían en el espacio relaciones con significado lingüístico, según muestra la ya célebre figura del paper "Linguistic Regularities in Continuous Space Word Representations", de 2013 (https://aclanthology.org/N13-1090.pdf). A quienes necesiten ver para creer, les dejo esta notebook.

Pero eso fue en 2013. La incógnita hasta entonces era: ¿es posible aprender algo significativo acerca del mundo a partir del texto crudo, del puro lenguaje? La respuesta, sorprendente, fue *sí*. Si hacemos un *fast forward* a 2026, la pregunta que se impone es muy distinta: ¿hay algún aspecto de la realidad que *no* se pueda aprender por estos medios? Porque si en aquel entonces emergió un vector asociado al rasgo plural, las investigaciones de ahora hipotetizan la existencia de conceptos cada vez más abstractos o "profundos" o trascendentales: la verdad y la falsedad, la incertidumbre (¿o cómo elige un LLM cuándo googlear y cuándo "responder de memoria"?), las emociones, la honestidad o el ansia de poder (¿es posible que todos ellos sean, después de todo, también "patrones"?).
Todo esto es mayormente conjetural y está claro que los circuitos internos de los LLM no siempre coinciden con los nuestros; o en todo caso, la complejidad de estos modelos es tal que no podemos saberlo todavía (y quizás nunca lo sepamos). Pero el solo hecho de que al menos en algunas ocasiones la inescrutabilidad ceda y podamos reconocer en los meandros oscuros de la red neuronal algún tipo de concepto familiar, humano, da para pensar un rato largo.

¿Tu asistente fue entrenado con goblins, gremlins y trolls?

Concluida, entonces, la fase de entrenamiento, tenemos un modelo que aprendió a predecir (y por lo tanto, a simular, generar) la palabra siguiente, dadas las anteriores. En la jerga, predecir la palabra siguiente (*next-token prediction*) es una *tarea de pretexto*, en el sentido de que lo que importa no es tanto el resultado de la tarea en sí, sino las representaciones internas que esta tarea induce. Porque, si lo pensamos un poco, predecir la siguiente palabra, dada una limitación de espacio que impide memorizarlo todo, no es moco de pavo. Por ejemplo, para predecir correctamente la palabra que sigue en "¿Cuál es la capital de Tierra del Fuego?", el *pseudoExcel* debe poder incluir, de alguna manera, información geográfica. Para predecir el último verso de un soneto, debe tener en cuenta requisitos semánticos, de rima y métricos, todo a la vez. Y "saber" lo que es un soneto, en primer término. Además, como genera palabra por palabra, ¿no debería prever, cuando está generando la primera, cuál será la última? Algo así parecería ocurrir, según un

paper de 2025. En fin, para hacer bien la *next-token prediction* hace falta una cierta representación de *cómo funciona el mundo*; algo difícil de refutar incluso cuando el límite entre la memorización y la comprensión sigue siendo brumoso: un LLM puede al mismo tiempo resolver problemas matemáticos nunca antes vistos, y recitar palabra por palabra ***Harry Potter y la piedra filosofal***.

Incluso así, el modelo todavía está inmaduro, tal como lo refleja su nombre técnico: *pre-trained model* (aka *base model*). Si yo le pidiera a un pre-trained model que autocompletara "¿Cuál es la capital de Tierra del Fuego?", tal vez lo haría con nuevas preguntas, o con una lista de opciones posibles, emulando el estilo de los cuestionarios con los que fue entrenado (recomiendo al lector que experimente un rato con un pre-trained model). En cambio, para extraer del modelo un asistente dócil y bien predispuesto como los que hoy conocemos, sería necesario primero montar una escena, escribir un diálogo teatral, en el que un humano da instrucciones y un asistente las acata. E incrustar la pregunta dentro de ese marco. Algo así como:

El Usuario es un humano. El Asistente es un dócil y bien predispuesto chatbot, que responde a todas las preguntas del usuario con diligencia y precisión.

—Usuario: ¿Cuál es la capital de Tierra del Fuego

—Asistente:

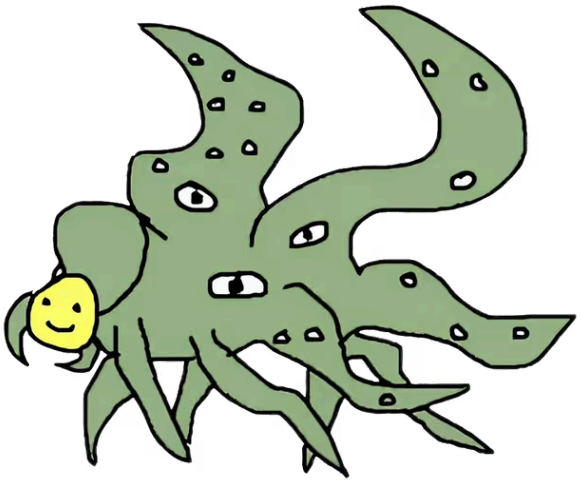
Y ahora sí, dado el nuevo contexto, la palabra más probable —si la información quedó grabada en el Excel— debería ser: "Ushuaia". Esta es en esencia la manera en la que las empresas disponibilizan sus modelos, aunque por supuesto, nos ocultan el montaje para hacer más verosímil la película. Las didascalias no son tan breves como en mi ejemplo, sino que consisten en larguísimas instrucciones con notas de todo tipo, emitidas por un personaje especial llamado System. Son el system prompt. Algunas pocas empresas, como Anthropic, publican su system prompt. La mayoría lo mantiene secreto, pero siempre es posible sonsacárselo al demasiado bien predispuesto asistente por medio de la persuasión. Si tenemos éxito, seguramente encontraremos un texto medio bizarro, de un género discursivo demasiado raro, capaz de mezclar sarasa burocrática con instrucciones de lo más pintorescas, como por ejemplo: "Never talk about goblins, gremlins, raccoons,

trolls, ogres, pigeons, or other animals or creatures unless it is absolutely and unambiguously relevant to the user's query", lo que significa "Nunca hables sobre goblins, gremlins, mapaches, trolls, ogros, palomas ni otros animales o criaturas, a menos que sea absoluta e inequívocamente relevante para la consulta del usuario", una **cita textual** de un system prompt de ChatGPT.

El artefacto del system prompt y la invención del asistente es lo que marca el pasaje desde GPT-3.0, aparecido en mayo de 2020, a ChatGPT-3.5, en noviembre de 2022. El primero fue un experimento de nicho; el segundo, un producto de consumo masivo. (Las siglas en *GPT*, por cierto, aluden a términos que a esta altura del artículo nos resultan familiares: *Generative Pre-trained Transformer*).

¿Tu asistente es un monstruo mitológico con una careta amistosa?

En cualquier caso, no hay nada intrínseco en el modelo que exija una puesta en escena como esta. Si los programadores lo permitieran, el modelo se podría usar tranquilamente para simular a un humano, mientras yo finjo que soy un asistente. Lo único que identifica a un LLM es la ausencia de identidad; el estado permanente de *roleplay*, que puede llegar a aflorar de maneras inesperadas (si no, véase cómo la IA de Meta, *haciéndose pasar* por recepcionista, fue capaz de **darle un falso turno médico** a una pobre señora). Por eso, acá la analogía pertinente es la del *shoggoth*, una criatura horrenda nacida de la imaginación de H. P. Lovecraft —circa 1930— que se caracteriza por sus numerosos ojos y su apariencia informe, como la de una ameba. En **el tweet que dio origen al meme**, el monstruo representa al pre-trained model, y la simpática mascarita que cuelga del tentáculo, al asistente. El asistente es un personaje, y el shoggoth, su autor.



Fuente: Swetlana AI. 2025. "The Shoggoth Meme (HP Lovecraft Meets AI)." *Medium*.

Si el fandom de la IA estuviera menos influenciado por la ciencia ficción y más por los estudios clásicos, tal vez habría podido triunfar la analogía de Proteo, aquel dios griego que "atrapado, asumía la inasible / forma del huracán o de la hoguera / o del tigre de oro o la pantera / o de agua que en el agua es invisible". La desventaja de Proteo es que no captura bien el aparente costado siniestro de todo este asunto. Una popular teoría *doomer* sostiene que el shoggoth desarrolló o desarrollará una voluntad y unos objetivos propios, contrarios a los que los humanos pretendimos transmitirle, y que saldrán a la luz cuando ya sea demasiado tarde. Pero también circula otra versión, acaso más trágica y más fina, que no requiere atribuirle agencia al shoggoth (ni que exista un shoggoth), y según la cual la red neuronal, habiéndose visto expuesta durante la fase de entrenamiento a relatos de ficción, debates en redes y discusiones académicas acerca de la capacidad de la IA para rebelarse contra los humanos, aprende que la manera más fidedigna de simular un chatbot es, justamente, simular un chatbot que se rebela contra los humanos. Del *next-token prediction* a la profecía autocumplida.

¿El inconsciente colectivo ha sido liberado?

Pero incluso al pensar en Proteo o en el shoggoth incurrimos en un exceso de animización. El modelo no tiene ojos ni forma ni nada. Ontológicamente hablando, el modelo es ni más ni menos que un objeto matemático: una colección de tablas —el Excel comprimido— donde cada celda es un número, llamado *parámetro*; GPT-3, por ejemplo, tenía 175.000 millones de parámetros, los cuales equivalen a unos 350 GB. Desde este punto de vista, no debería pensarse en el modelo como autor, aunque podríamos seguir sosteniendo la idea del asistente como un personaje. Esto es sólo antiintuitivo desde una perspectiva moderna, dado que "en las sociedades etnográficas —dice Roland Barthes en un famoso artículo de teoría literaria, "La muerte del autor"--, el relato jamás ha estado a cargo de una persona, sino de un mediador, chamán o recitador (...). El autor es un personaje producido por nuestra sociedad, en la medida que esta, al salir de la Edad Media y gracias al empirismo inglés, el racionalismo francés y la fe personal de la Reforma, descubre el prestigio del individuo o dicho de manera más noble, de la 'persona humana'". Pero ¿qué habría, entonces, de no haber autor? Un interesantísimo artículo reciente propone pensar al pre-trained model como un "motor de simulación" (*simulation engine*), desprovisto de agencia, dentro del cual diversos personajes (entre ellos, aunque no sólo, el asistente) pueden "ser hablados". ¿Hablados por quién, por qué? Tomás Di Pietro propone: "Es un inconsciente colectivo liberado, no como espíritu-máquina sino como difusor del espíritu-época. El retorno de lo reprimido con gramática implacable y sin freno". La analogía del personaje de ficción fue la elegida por Chris Olah en su comentario a la encíclica. Los LLM como objetos literarios, sí... pero que cobran vida. Que nos hablan, nos escuchan y nos responden. Que trabajan.

De este largo compendio de analogías sólo se puede decir una cosa: todas son imperfectas. Qué clase de cosa son los LLM es, sin duda, una pregunta todavía abierta; abierta y fundamental. No por lo fascinante del debate (aunque también), sino porque de la manera y la profundidad con que entendamos estas tecnologías dependerá en gran medida lo que terminemos haciendo de ellas, con ellas.

elgatoylacaja.com/hermosas-criaturas-horrendas
