ENSAYO 7

Calibración e IA: Un paseo en el carro de Faetón

Claudia Marsico · Hernán Inverso

¿Y si el problema está en cómo lo dice? Una exploración urgente de modelos de lenguaje desencajados.

Gustave Moreau, La caída de Faetón (1878)

Un experto sabe muchas cosas, pero cuando se acerca al borde, baja la voz o titubea. Un LLM puede mantener el tono cuando responde 1815 al año de Waterloo o lanza un dato dudoso como resultado de un cálculo en el que suele acertar poco. El fenómeno suele llamarse “ausencia de calibración” y atañe al grado de seguridad en la respuesta. La confianza debería subir y bajar con la evidencia. Ay de nosotros cuando no sucede.

En línea con los desastres que acarrea, el nombre se mezcla con armas y explosiones. Allá por el siglo XVI en Francia "calibre" era el diámetro de la boca de un cañón, tal vez del árabe *qālib*, "molde", a su vez del griego *kalópous*, la horma del zapatero, o más probablemente del latín medieval *qua libra*, "de qué peso". Como sea, el verbo "calibrar" se encuentra en francés en el s. XVII como tomar la medida de algo y se expande pronto a otras lenguas. La acepción que nos importa, graduar un instrumento para que mida sin error, es del s. XIX.

Pocos temas se entrelazan tanto con la historia de Occidente como este. Meden agan, nada en exceso, es la fórmula délfica de la calibración. El mito está lleno de ejemplos de los horrores que conlleva ignorarla. Tomemos el caso de Faetón. Hijo de Helios, el Sol, para presumir de su origen, suplica conducir el carro de su padre. Helios se horroriza y trata de disuadirlo, explicando que el camino es empinado y los caballos escupen fuego. Ni Zeus, le dice, en la versión de Ovidio en *Metamorfosis*, se anima a manejarlo. Faetón se obstina y toma las riendas. Inmediatamente los caballos desconocen la guía y se desbocan. El carro asciende locamente y quema el cielo, baja en picada y quema la tierra, seca los ríos, abre los desiertos de Libia y tuesta para siempre la piel de los pueblos del sur. Para frenar el incendio del mundo, Zeus no tiene más remedio que usar el rayo y Faetón cae ardiendo. Calibró mal su habilidad de conducir el carro y su seguridad infundada lo mató.

Gustave Moreau, La caída de Faetón (1878)

En la fábula de Esopo, el pastorcito que grita "¡lobo!" por diversión cuando no hay ninguno entrena a los vecinos en la desconfianza. Cuando el lobo aparece de verdad, sus gritos ya no valen nada, porque gastó su crédito en falsas alarmas. El pastorcito falla tanto que no le creen cuando acierta. También trata sobre la calibración *Otelo*, de Shakespeare, escrita alrededor de 1603. Cuando el moro de Venecia pide una "prueba ocular" de la traición de Desdémona, Yago le entrega apenas un pañuelo y unas insinuaciones. Con ese material mínimo Otelo la mata. Tan poco es lo que sabe y tanto lo que cree saber. En la misma estirpe de quienes fallan al graduar está el Quijote que ve gigantes donde hay molinos y el rey Lear que mide el amor de sus hijas por el tamaño de sus discursos.

¿De qué otra cosa que de calibración se ocupa Sócrates cuando define su sabiduría por no creerse sabio? Dos mil años más tarde Montaigne haría de ese temple una divisa, "Que sais-je?", qué sé yo, que mandó grabar en una medalla. El término de oro de la calibración es "probable", acuñado por Cicerón en su *Academica* (45 a. C.) para traducir el griego *pithanón*, "lo creíble", el criterio con que el escéptico Carneades proponía moverse por un mundo donde la certeza se nos escabulle y los estoicos hicieron descansar toda su epistemología en el asentimiento que damos a las impresiones, de modo tal que su sabio infalible, en sintonía con el cosmos, es un calibrador que no se equivoca.

Ese asentimiento se volvió un arte en la modernidad. Para Descartes, en la cuarta de sus *Meditaciones metafísicas* (1641), estamos a merced de una voluntad más amplia que el entendimiento y de claridad limitada que nos precipita en el error, es decir, nos descalibramos al afirmar con más fuerza de la autorizada por la evidencia. Poco después Jacob Bernoulli, en su *Ars Conjectandi* (1713), definió la probabilidad como un grado de certeza y demostró que, con suficientes intentos, la frecuencia de un hecho se acerca a su probabilidad, tendiendo el puente entre la creencia y la medida. Por su parte, el

obispo Joseph Butler, en *La analogía de la religión* (1736), acuñó la frase "la probabilidad es la guía misma de la vida", porque la evidencia probable admite grados, desde la certeza moral más alta hasta la presunción más baja.

John Locke le dedicó, en su *Ensayo sobre el entendimiento humano* (1689), un capítulo a los "grados del asentimiento", donde pedía creer cada cosa con una firmeza proporcional a sus indicios. David Hume lo sintetizó, en su *Investigación sobre el entendimiento humano* (1748), en la idea de que un hombre sabio gradúa su creencia según la evidencia. El matemático William Clifford lo llevó a su forma más dura en "La ética de la creencia" (1877), diciendo que está mal, siempre, en todas partes y para cualquiera, creer algo sobre la base de una evidencia insuficiente. William James miró entre las grietas y le respondió en "La voluntad de creer" (1896) que hay decisiones vitales que no admiten esperar a la evidencia completa. A veces simplemente hay que animarse a apostar. De todos modos, en lo que a nuestro tema respecta, los dos coinciden en la importancia de saber de dónde viene nuestra seguridad y no andar por ahí convencidos de algo en lo que tenemos pocas posibilidades de acertar.

El mandamiento de la probabilidad, enunciado por Dennis Lindley en *Making Decisions* (1971), dicta que nunca debemos asignar una certeza del cero ni del cien por ciento, salvo a las verdades lógicas, porque los extremos clausuran el aprendizaje de indicios futuros. Se la llama "regla de Oliver Cromwell", porque Lindley se inspiró en la súplica que en 1650 hizo a los presbiterianos escoceses: "les ruego, por las entrañas de Cristo, que consideren que pueden estar equivocados". Afirmar cada cosa con la misma seguridad destruye el margen para equivocarse, cosa que hacemos a menudo. Según Sarah Lichtenstein y Baruch Fischhoff (1977), la gente segura al setenta por ciento acierta cerca de la mitad de las veces, mientras que la confianza extrema es la peor calibrada. En 1999 David Dunning y Justin Kruger desplegaron la otra cara de la moneda al mostrar casos donde los menos competentes son los que peor miden su propia competencia, faltos como están de criterio de medida.

La meteorología, ese gran desafío cotidiano, dio un paso adelante cuando en 1950 Glenn Brier inventó un sistema para puntuar los pronósticos del tiempo que todavía lleva su nombre, asociando la calibración con la correlación entre confianza y frecuencia del acierto. Un pronosticador honesto no acierta siempre, pero sabe cuánto suele acertar y lo dice. En las décadas siguientes las mediciones ganaron en precisión y mejoró el llamado "diagrama de fiabilidad", que enfrenta en un gráfico la seguridad que el sistema declara y el porcentaje que acierta. Si dice noventa y acierta noventa, cae sobre la diagonal, la línea de honestidad; si dice noventa y acierta setenta, la curva se hunde por debajo y esa distancia hasta la diagonal, promediada a lo largo de toda la escala se llama error de calibración. Sirve para saber si el "estoy seguro" de la máquina vale algo y por eso se usó en redes neuronales que clasifican imágenes, bastante antes de que existieran los chatbots. Sigue siendo hoy el instrumento estándar de la calibración para los modelos de lenguaje.

Las redes neuronales se volvieron más precisas y, al mismo tiempo, más sobreconfiadas (Guo et al. 2017). Las de los años noventa, pequeñas, estaban bien calibradas; las recientes, profundas, aciertan más pero exageran su certeza, y las mismas decisiones de diseño que subieron la precisión (más capas, más parámetros, la normalización por lotes), hundieron la calibración. Sin embargo, alcanza con una

sola variable, la temperatura, que ablanda o endurece de golpe todas las probabilidades del modelo, para recalibrar buena parte del desajuste. Si una sola perilla arregla tanto, la sobreconfianza es un sesgo parejo de toda la escala.

Con los modelos de lenguaje, el mecanismo es claro. Un modelo base aprende ajustando sus predicciones a la frecuencia con que cada palabra aparece en los datos: cuanto más se aparta de esas proporciones, mayor es la penalización. Por eso, en principio, sus probabilidades reflejan bastante bien lo que efectivamente ocurre en el material con que fue entrenado. En el informe técnico de GPT-4 (2023), la curva de fiabilidad del modelo base es casi una diagonal perfecta: cuando declara mayor confianza, también acierta en una proporción equivalente. Después llega el ajuste que lo convierte en asistente y cambia el objetivo. La retroalimentación humana prefiere lo que suena seguro, de modo que la política óptima se traga la duda y con ella la calibración. En el mismo informe de GPT-4, la diagonal que traía el modelo base se arquea después del ajuste y la aguja se pega arriba.

La historia no es fatal, porque la señal calibrada no desaparece: queda enterrada. Un modelo grande, interrogado del modo adecuado, puede estimar la probabilidad de que su propia respuesta sea correcta; esa capacidad mejora con el tamaño y el ajuste por refuerzo la debilita sin borrarla (Kadavath 2022). Un modelo puede aprender a decir en lenguaje natural cuán seguro está, incluso sin acceder a sus probabilidades internas (Lin et al. 2022) y en los modelos alineados la confianza expresada con palabras suele estar mejor calibrada que la que se obtiene de sus estados internos y reduce cerca de la mitad del error (Tian 2023). El conocimiento de la propia ignorancia seguía allí, archivado en otro cajón. Sin embargo, preguntarle al modelo no obra milagros. Al poner su confianza en palabras, el modelo también tiende a exagerarla, concentrando sus respuestas entre el noventa y el cien por ciento, como si imitara la fanfarronería humana (Xiong 2024). La calibración puede recuperarse, entonces, pero solo a medias.

¿Se podría entrenar a un modelo en la actitud socrática, el emblema de la duda bien medida? En parte, eso es lo que buscan los modelos de razonamiento o Thinking. Antes de responder, descomponen el problema, ensayan caminos, revisan sus pasos y tratan de detectar sus propios errores. Ese tiempo adicional suele mejorar el acierto. También funcionan los procedimientos dialécticos a través de consejos de modelos. Sin embargo, razonar más no equivale a estar mejor calibrado. La revisión puede llevar a defender un error con mayor habilidad. Una capa socrática añadida puede producir apenas duda representada: preguntas, reservas y cautelas que envuelven la misma sobreconfianza. Una cautela indiscriminada tampoco sirve. El modelo que responde siempre “cincuenta por ciento” no distingue lo seguro de lo incierto. Calibrar es variar la duda según la dificultad del terreno.

El procedimiento socrático puede ayudar a producir mejores razones y datos para la evaluación, pero no constituye por sí solo una cura. Y tiene sentido. El modelo puede aprender a reconocer patrones asociados con su propio error y a decir “no lo sé” en los casos adecuados, lo cual implica una conducta epistémicamente prudente, pero no hay nadie intentando conocerse a sí mismo. El modelo no tiene una historia de adquisición, un registro interno de cuándo y con qué firmeza aprendió cada cosa, así que todo lo que dice sale con el mismo peso, lo comprobado mil veces y lo hilado recién, sin una

etiqueta que los separe. Calibrarse supone un yo que se juega en acertar, que puede estar equivocado y pagarlo y que por eso aprende a cuidar su certeza, dejándose siempre, como pedía Cromwell, un margen para estar equivocado.

Faetón se descalibró y ardió. Su tragedia tiene, al menos, la forma de una tragedia. Hubo un padre que le rogó que no lo hiciera, una caída, un cuerpo, un castigo que convirtió el exceso en lección para el resto. El modelo toma las riendas del carro del Sol con la misma ceguera de Faetón, pero no siente el vértigo, no se le queman las manos, no cae. Es un Faetón sin caída, un exceso al que no le sigue ningún desastre, porque no hay nadie ahí que pueda desastrarse. Lo que puede arder es la tierra de abajo, donde estamos nosotros, los que recibimos el dato falso dicho en el tono seguro. El instrumento que gradúa cuánto merece creerse lo que nos dicen no está en el carro. Está, como siempre, de nuestro lado y nos toca no soltarlo.

◊

◊

Mapas del laberinto es una serie de Phantom Maze, AI & Language Lab. La colección vive en **phantommaze.cloud**; los ensayos aparecen también en phantommazeailabes.substack.com. La suscripción es gratuita.

Algunas de las ideas que enmarcan este texto se desarrollan en *El fantasma sin ego. Un mapa en el laberinto de la inteligencia artificial*, de Hernán Inverso y Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-10-2).

PHANTOM MAZE

Leer este ensayo en línea · Todos los ensayos · El libro

G. W. Brier, “Verification of Forecasts Expressed in Terms of Probability” (*Monthly Weather Review*, 1950). S. Lichtenstein, B. Fischhoff y L. Phillips, “Calibration of Probabilities: The State of the Art” (1977/1982); D. Dunning y J. Kruger, “Unskilled and Unaware of It” (1999). C. Guo, G. Pleiss, Y. Sun y K. Q. Weinberger, “On Calibration of Modern Neural Networks” (ICML 2017; OpenAI, *GPT-4 Technical Report* (2023, la calibración del modelo base degradada tras el ajuste, fig. 8); A. T. Kalai y col. (OpenAI), “Why Language Models Hallucinate” (2025); S. Kadavath y col. (Anthropic), “Language Models (Mostly) Know What They Know” (2022); S. Lin, J. Hilton y O. Evans, “Teaching Models to Express Their Uncertainty in Words” (2022); K. Tian, E. Mitchell y col., “Just Ask for Calibration” (EMNLP 2023); M. Xiong y col., “Can LLMs Express Their Uncertainty?” (ICLR 2024); S. Farquhar, J. Kossen, L. Kuhn e Y. Gal, “Detecting Hallucinations in Large Language Models Using Semantic Entropy” (*Nature*, 2024); “Mind the Confidence Gap: Overconfidence, Calibration, and Distractor Effects in Large Language Models” (2025).