MAPS OF THE LABYRINTH

A SERIES FROM PHANTOM MAZE — AI & LANGUAGE LAB

ESSAY 7

Calibration and AI: A Ride in Phaethon's Chariot

Claudia Marsico · Hernán Inverso

What if the problem lies in how it says it? An urgent look at language models that have come unhinged.

Gustave Moreau, The Fall of Phaethon (1878)

An expert knows a great many things, but when he nears the edge he lowers his voice or hesitates. A language model can hold the same tone whether it answers 1815 to the year of Waterloo or floats a doubtful figure as the result of a calculation it rarely gets right. The phenomenon is usually called a lack of calibration, and it concerns the degree of confidence carried by an answer. Confidence ought to rise and fall with the evidence. Woe to us when it does not.

In keeping with the disasters it brings, the word keeps company with weapons and explosions. Back in sixteenth-century France, *calibre* was the diameter of a cannon's mouth, perhaps from the Arabic *qālib*, "mould", itself from the Greek *kalópous*, the shoemaker's last, or more likely from the medieval Latin *qua libra*, "of what weight". Either way, the verb "to calibrate" turns up in seventeenth-century French meaning to take the measure of something, and it soon spreads to other languages. The sense that matters to us, adjusting an instrument so that it measures without error, is nineteenth-century.

Few subjects are as tightly braided into the history of the West as this one. *Mēdèn ágan*, nothing in excess, is the Delphic formula for calibration. Myth is full of examples of the horrors that follow from ignoring it. Take the case of Phaethon. Son of Helios, the Sun, and eager to flaunt his lineage, he begs to drive his father's chariot. Helios is appalled and tries to talk him out of it, explaining that the road is steep and the horses breathe fire. Not even Zeus, he tells him in Ovid's version in the *Metamorphoses*, dares to handle it. Phaethon insists and takes the reins. At once the horses no longer recognise the hand that guides them and bolt. The chariot climbs wildly and burns the sky, plunges and burns the earth, dries the rivers, opens the deserts of Libya and scorches for ever the skin of the peoples of the south. To stop the world from burning, Zeus has no choice but the thunderbolt, and Phaethon falls in flames. He miscalibrated his skill at driving the chariot, and unfounded confidence killed him.

In Aesop's fable, the shepherd boy who cries "wolf!" for fun when there is none trains his neighbours in distrust. When the wolf finally comes, his cries are worth nothing, because he spent his credit on false alarms. The boy is wrong so often that no one believes him when he is right. Shakespeare's *Othello*, written around 1603, is also about calibration. When the Moor of Venice demands "ocular proof" of Desdemona's betrayal, Iago hands him barely a handkerchief and a few insinuations. On that minimal material Othello kills her. So little is what he knows and so much what he believes he knows. Of the same lineage of those who fail at grading are the Quixote who sees giants where there are windmills and the King Lear who measures his daughters' love by the length of their speeches.

What else is Socrates concerned with, if not calibration, when he defines his own wisdom by not believing himself wise? Two thousand years later Montaigne would make that temper a device, "*Que sais-je?*", what do I know, and had it struck on a medal. The golden term of calibration is "probable", coined by Cicero in his *Academica* (45 BC) to translate the Greek *pithanón*, "the credible", the criterion with which the sceptic Carneades proposed to move through a world where certainty slips away from us; and the Stoics rested their whole epistemology on the assent we give to impressions, so that their infallible sage, in tune with the cosmos, is a calibrator who never errs.

That assent became an art in the modern age. For Descartes, in the fourth of his *Meditations on First Philosophy* (1641), we are at the mercy of a will broader than the understanding and of limited clarity, which hurls us into error; we lose our calibration, that is, by affirming with more force than the evidence authorises. Shortly afterwards Jacob Bernoulli, in his *Ars Conjectandi* (1713), defined probability as a degree of certainty and showed that, given enough trials, the frequency of an event approaches its probability, throwing the bridge between belief and measurement. For his part Bishop Joseph Butler, in *The Analogy of Religion* (1736), coined the phrase "probability is the very guide of

life", because probable evidence admits of degrees, from the highest moral certainty to the lowest presumption.

John Locke devoted a chapter of his *Essay Concerning Human Understanding* (1689) to the "degrees of assent", where he asked that each thing be believed with a firmness proportioned to its grounds. David Hume condensed it, in his *Enquiry Concerning Human Understanding* (1748), into the idea that a wise man proportions his belief to the evidence. The mathematician William Clifford took it to its hardest form in "The Ethics of Belief" (1877), saying that it is wrong always, everywhere, and for anyone, to believe anything upon insufficient evidence. William James looked into the cracks and answered him in "The Will to Believe" (1896) that there are vital decisions that cannot wait for the evidence to be complete. Sometimes one simply has to dare to bet. In any case, as far as our subject goes, the two agree on how much it matters to know where our confidence comes from, and not to walk around convinced of something we have little chance of getting right.

The commandment of probability, stated by Dennis Lindley in *Making Decisions* (1971), says that we must never assign a certainty of zero or of one hundred per cent, except to logical truths, because the extremes shut down any learning from future evidence. It is known as Cromwell's rule, because Lindley took his cue from the plea Oliver Cromwell made to the Scottish Presbyterians in 1650: "I beseech you, in the bowels of Christ, think it possible that you may be mistaken." Asserting everything with the same confidence destroys the margin for being wrong, which we often are. According to Sarah Lichtenstein and Baruch Fischhoff (1977), people who are seventy per cent sure are right about half the time, while extreme confidence is the worst calibrated of all. In 1999 David Dunning and Justin Kruger unfolded the other side of the coin by showing cases where the least competent are the worst at gauging their own competence, lacking as they do the very yardstick.

Meteorology, that great daily challenge, took a step forward when in 1950 Glenn Brier invented a way of scoring weather forecasts that still bears his name, tying calibration to the match between stated confidence and the frequency of being right. An honest forecaster is not always right, but knows how often he tends to be right and says so. In the following decades the measurements gained precision and the so-called reliability diagram improved, plotting against each other the confidence the system declares and the percentage it gets right. If it says ninety and is right ninety, it falls on the diagonal, the line of truthfulness; if it says ninety and is right seventy, the curve sinks below it, and that distance from the diagonal, averaged across the whole scale, is called calibration error. It serves to tell whether the machine's "I am sure" is worth anything, and for that reason it was used on neural networks that classify images, well before chatbots existed. It remains today the standard instrument of calibration for language models.

Neural networks became more accurate and, at the same time, more overconfident (Guo et al. 2017). The small ones of the nineties were well calibrated; the recent deep ones are right more often but overstate their certainty, and the very design decisions that raised accuracy (more layers, more parameters, batch normalisation) sank calibration. Yet a single variable, temperature, which softens or

stiffens all the model's probabilities at a stroke, is enough to recalibrate much of the mismatch. If one knob fixes that much, overconfidence is a bias spread evenly across the whole scale.

With language models the mechanism is clear. A base model learns by fitting its predictions to the frequency with which each word appears in the data: the further it strays from those proportions, the greater the penalty. That is why, in principle, its probabilities reflect fairly well what actually occurs in the material it was trained on. In the GPT-4 technical report (2023), the base model's reliability curve is almost a perfect diagonal: when it declares greater confidence, it is right in an equivalent proportion. Then comes the tuning that turns it into an assistant, and the objective changes. Human feedback prefers what sounds confident, so the optimal policy swallows doubt and, with it, calibration. In that same GPT-4 report, the diagonal the base model brought with it arches after tuning and the needle sticks at the top.

The story is not fated, because the calibrated signal does not disappear: it is buried. A large model, questioned in the right way, can estimate the probability that its own answer is correct; that ability improves with size, and reinforcement tuning weakens it without erasing it (Kadavath 2022). A model can learn to say in plain language how confident it is, even without access to its internal probabilities (Lin et al. 2022), and in aligned models the confidence expressed in words is usually better calibrated than what can be read off their internal states, cutting close to half the error (Tian 2023). The knowledge of its own ignorance was still there, filed in another drawer. Asking the model, however, works no miracles. In putting its confidence into words, the model also tends to inflate it, crowding its answers between ninety and one hundred per cent, as if it were imitating human bluster (Xiong 2024). Calibration can be recovered, then, but only halfway.

Could a model be trained in the Socratic attitude, the emblem of well-measured doubt? In part that is what reasoning or Thinking models are after. Before answering they break the problem down, try out paths, review their steps and attempt to catch their own mistakes. That extra time usually improves accuracy. Dialectical procedures through councils of models also work. Yet reasoning more is not the same as being better calibrated. Review can lead to defending an error with greater skill. An added Socratic layer may produce no more than represented doubt: questions, reservations and cautions wrapped around the same overconfidence. Indiscriminate caution is no use either. The model that always answers "fifty per cent" does not tell the certain from the uncertain. To calibrate is to vary one's doubt with the difficulty of the ground.

The Socratic procedure can help produce better reasons and better data for assessment, but it is no cure in itself. And that makes sense. The model can learn to recognise patterns associated with its own error and to say "I don't know" in the right cases, which amounts to epistemically prudent conduct, but there is no one there trying to know himself. The model has no history of acquisition, no internal record of when and how firmly it learned each thing, so everything it says comes out with the same weight, what has been checked a thousand times and what has just been spun, with no label to separate them. To calibrate oneself presupposes a self with something at stake in being right, one that

can be wrong and pay for it and that therefore learns to look after its certainty, always leaving itself, as Cromwell asked, a margin for being mistaken.

Phaethon lost his calibration and burned. His tragedy has, at least, the shape of a tragedy. There was a father who begged him not to do it, a fall, a body, a punishment that turned excess into a lesson for everyone else. The model takes the reins of the Sun's chariot with the same blindness as Phaethon, but it feels no vertigo, its hands do not burn, it does not fall. It is a Phaethon without a fall, an excess followed by no disaster, because there is no one there who can be undone. What can burn is the earth below, where we are, the ones who receive the false figure said in the confident tone. The instrument that grades how much what we are told deserves to be believed is not in the chariot. It is, as always, on our side, and it falls to us not to let go of it.

◊

Maps of the Labyrinth is a series from Phantom Maze, AI & Language Lab. The collection lives at **phantommaze.cloud**; new essays also appear at phantommazelab.substack.com. Subscription is free.

Some of the framing ideas in this piece are developed in *The Egoless Phantom: Mapping the AI Labyrinth*, by Hernán Inverso and Claudia Marsico (Phantom Maze Press, 2026; ISBN 978-987-3729-16-4).

PHANTOM MAZE

Read this essay online · All essays · The book

References. G. W. Brier, "Verification of Forecasts Expressed in Terms of Probability" (*Monthly Weather Review*, 1950). S. Lichtenstein, B. Fischhoff and L. Phillips, "Calibration of Probabilities: The State of the Art" (1977/1982); D. Dunning and J. Kruger, "Unskilled and Unaware of It" (1999). C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On Calibration of Modern Neural Networks" (ICML 2017); OpenAI, *GPT-4 Technical Report* (2023, the base model's calibration degraded after tuning, fig. 8); A. T. Kalai et al. (OpenAI), "Why Language Models Hallucinate" (2025); S. Kadavath et al. (Anthropic), "Language Models (Mostly) Know What They Know" (2022); S. Lin, J. Hilton and O. Evans, "Teaching Models to Express Their Uncertainty in Words" (2022); K. Tian, E. Mitchell et al., "Just Ask for Calibration" (EMNLP 2023); M. Xiong et al., "Can LLMs Express Their Uncertainty?" (ICLR 2024); S. Farquhar, J. Kossen, L. Kuhn and 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).