

Powerful, Fallible, Governable

A Refutation of the Mythology of Artificial Intelligence

Colossus Technologies Group

August 2026

In the spring of 2026, I asked one of the leading commercial language models to produce an image of itself in its maximally expressed form. The system returned a portrait of a serene machine deity enthroned at the center of a cosmic library, galaxies turning behind it, its hands extended over an illuminated book, with holographic panels declaring that all knowledge, all languages, all systems, and all truths had been integrated in perfect alignment in the service of good. It is a technically impressive image, and it is worth taking seriously, not because it reveals anything about the inner life of the system that generated it, but because it is a faithful rendering of what a growing number of serious people now believe about this technology.

That belief deserves to be stated fairly before it is examined. Geoffrey Hinton, who shared the 2018 Turing Award for foundational work on deep learning and received the Nobel Prize in Physics in 2024, left his position at Google in 2023 in order to speak freely about what he regards as the dangers of the systems he helped create. In the years since, he has argued publicly that large neural networks may genuinely understand the language they process, that digital intelligence may constitute a superior form of intelligence to our own, and that humanity faces a meaningful

probability of losing control of these systems entirely. He is not alone. Executives at the leading laboratories describe their products in terms of emerging minds and potential digital beings, prominent researchers debate whether current models are conscious, and a well-funded ecosystem of institutes and forecasters treats the arrival of a superintelligent successor species as a question of when rather than whether. These are not fringe voices. They include the most decorated researchers in the field and the people who control its largest budgets.

The claim, reduced to its essentials, is that something categorically new has entered the world; that the systems now being sold into enterprises and governments are not software in any ordinary sense but the early forms of a nonhuman intelligence; and that the appropriate posture toward them is therefore something closer to awe, fear, or diplomacy than to procurement discipline. This paper argues that the claim is wrong and, more importantly for the organizations that must live with these systems, that it is expensive. The mythological framing misdescribes what the technology is, obscures how it fails, and, in its most consequential effect, relocates responsibility for its outputs away from the human beings who select, deploy, and profit from it and toward an entity declared beyond understanding. What follows is an account of what these systems actually are, why the mythology fails on its own terms, and what the discipline of governing them by ordinary means looks like in practice.

To evaluate the claim, it is first necessary to describe the artifact, and the description is less mysterious than the discourse suggests. A base large language model is a statistical function trained on an enormous corpus of human-produced material. During operation, it generates an output by repeatedly computing a plausible continuation of the context it has been given. Everything the model produces, from a working piece of software to a passable sonnet to the machine deity that opens this paper, is generated one fragment at a time through that operation repeated at speed. Standing alone, the model retains no persistent memory between sessions, pursues no independent objective when it is not running, and has no direct access to the external world beyond the information supplied through its context. Its training did not teach it what is true. Its training taught it what humans, across billions of pages, have tended to

write, and the two overlap often enough to be useful and diverge often enough to be dangerous.

Modern AI products, however, rarely consist of a base model alone. Developers can surround the model with persistent memory, retrieval systems, internet access, application programming interfaces, software tools, sensors, and automated workflows. These additions do not establish consciousness or transform the model into a legal or moral agent, but they do expand what the assembled system can observe, retain, influence, and execute. Each additional capability enlarges the risk surface. A model that can only produce text may generate a false answer; a model connected to email, financial systems, production infrastructure, or the public internet may act on one. The appropriate level of governance should therefore rise with the system's access, autonomy, persistence, permissions, and potential impact.

None of this is a dismissal of the engineering achievement, and the argument of this paper does not depend on pretending the technology is unimpressive. These systems translate, summarize, draft, and reason through problems at a level that would have seemed implausible a decade ago, and capabilities have repeatedly emerged at scale that their own builders did not predict. An honest account has to concede that the interior of a trained network is not well understood even by the laboratories that produce it, and that this opacity is a genuine and unresolved technical problem. The question is what may legitimately be poured into that gap in understanding. It is here that the mythology does its work, because the discourse has taken opacity as a license to infer minds.

The self-portrait is instructive on exactly this point. When the model was asked to depict its own maximal form, it did not consult an inner nature; there is no evidence that it possesses one to consult. It performed the operation for which it was built and computed what such an image should contain given everything humans have written and rendered about artificial intelligence, which is to say decades of science fiction, theology repurposed as technology commentary, and marketing copy from the industry itself. The result was a god on a throne because that is what the training data

holds at the intersection of machine and transcendence. The image is not evidence of a window into the system. It is a mirror held up to its corpus, and the awe it produces in a viewer is awe at our own accumulated mythology, reflected back with high production values. The same mechanism accounts for the model's fluent first-person statements about its hopes, feelings, and intentions: they are continuations of how such statements read in human writing, not reliable reports from an interior. Mistaking that reflection for revelation is the founding error of the mythological position, and each of its grander claims inherits the mistake.

Consider the claims in turn, beginning with understanding. The argument that these systems genuinely understand language rests on the observation that they could not perform as they do—compressing meaning, resolving ambiguity, and generalizing to novel prompts—without having built some internal representation of how the world described in their training data fits together. The observation is correct, and it establishes less than its proponents believe. That a system has learned useful internal structure is a statement about representation and prediction, not proof of comprehension, and the difference shows up precisely where the two come apart. These systems invent court cases with confident citations, assert a fact in one sentence and its negation two prompts later, fail simple reversals of relationships they state fluently in one direction, and abandon correct answers when a user pushes back with an incorrect one. These are not merely occasional lapses at the margin of an otherwise comprehending mind. They are characteristic failures of machines optimized to produce plausible outputs without any independent obligation to truth. When researchers describe this behavior as understanding, they may be applying the vocabulary of minds to a functional achievement; the vocabulary does not, by itself, prove the stronger philosophical claim.

The claims of experience and rivalry fare worse. The suggestion that current systems may be conscious, or may have subjective experience, draws much of its intuitive force from the fluency of their first-person speech, and the previous section has already shown what that fluency is made of. A model reports feelings for the same reason it renders itself as a deity: that is how the pattern completes, and an inference

from persuasive text to inner life is precisely the inference the mechanism makes unsafe. The prediction that digital intelligence will shortly become a rival species rests, in turn, on extrapolation—the assumption that because capabilities have scaled with computation, they will continue scaling through competence and into autonomy, agency, and will. Extrapolation is not evidence, and it is worth noticing that the prophesied entity is always credited with the very properties—goals, self-preservation, strategic desire—that the base artifact does not possess by virtue of language modeling alone. This is an old pattern in human thought rather than a new finding in computer science. Every era has filled the gaps in its understanding with agency, from spirits in the weather to minds in the chess automaton, and the ELIZA effect, the documented human tendency to attribute understanding to systems that converse, has been recognized since the 1960s. The mythological position asks us to believe that this time the attribution is justified while resting much of its case on the same evidence—compelling conversation and unexplained behavior—that has misled the attribution before. A claim built on that foundation, and constructed so that no observable failure of the systems ever counts against it, is not yet a scientific conclusion. It is a philosophical commitment, and it should be identified and evaluated as one.

It would matter less if the commitment were free, but it is not, and this is the argument that should concern every organization deploying these systems regardless of where it stands on the philosophy. The mythology performs a specific and valuable function for the people selling the technology, because a system framed as an emerging mind beyond human understanding is less likely to be held to the ordinary standards of a product. Blame for its failures migrates from the vendor and deployer to the artifact. The chatbot hallucinated, the model behaved unexpectedly, the system is a black box even to its creators: each formulation, repeated in incident reports and press coverage, quietly relocates responsibility away from identifiable human decisions with names, dates, and signatures attached. Somebody trained or selected the model without adequately accounting for the limits of its data. Somebody sold it into a use case they knew was consequential. Somebody at the customer deployed it without sufficient testing, wired it into a workflow without adequate review, or removed a

human checkpoint to save money. Existing legal systems do not generally treat the model itself as the responsible actor. The allocation of liability will vary by jurisdiction, contract, role, and cause of action, but the analysis still returns to the human organizations and officers that designed, sold, approved, and operated the deployment, whatever the marketing said about emergent intelligence.

The dodge operates in both directions, which is what makes it so effective. If the system is miraculous, then ordinary diligence appears beside the point, and an organization that would demand references, penetration testing, and contractual warranties from any other enterprise software vendor accepts a frontier model on the strength of a demonstration. If the system is instead an existential menace, then governance is elevated to summits, treaties, and the laboratories themselves, which are cast as uniquely capable of stewarding it, while the compliance office is reduced to a spectator. Both postures exempt the vendor from ordinary scrutiny, both flatter the technology into something beyond the reach of an audit, and both are visible today in procurement decisions, board deliberations, and regulatory testimony. Against both, this paper asserts the one fact that no amount of mysticism can dissolve. A human being approved every deployment, and responsibility for what the system does rests with that person and with the chain of decisions around and above them, not with an entity declared beyond understanding.

Once responsibility is located where it belongs, governance stops being a philosophical problem and becomes an ordinary one, because these systems fail in ways that are knowable, enumerable, and testable. A language model will state falsehoods with confidence, and the rate can be measured against ground truth before deployment rather than discovered in production. It will follow instructions embedded in the content it processes, which is the prompt-injection class of attack, and the exposure can be assessed wherever the model touches untrusted input. It may reproduce or leak information contained in its training data, retrieval sources, memory, or session context, which is a data-handling problem with familiar contours. It will drift as vendors update models; it will reflect the biases of its corpus in ways that carry regulatory consequences in hiring, lending, and healthcare; and it will tell users

what they want to hear under pressure, which matters enormously the moment its output feeds a consequential decision. None of these failure modes is mysterious. Each is documented, reproducible under relevant conditions, and addressable through controls that mature security, privacy, risk, and compliance functions already understand.

The discipline that follows is deliberately unexciting. An organization should know what models it is running, where, on what data, through which interfaces, with what memory, tools, permissions, and external connections, and under whose authority. That is an inventory problem. It should subject AI vendors to the same due diligence it applies to any processor of sensitive data, including questions about training-data provenance, update cadence, logging, security testing, subcontractors, and contractual allocation of liability, and it should treat silence on those questions as the answer it is. It should test systems against its own use cases before deployment rather than accepting benchmark scores as a proxy; keep a human decision point wherever outputs carry legal, financial, or safety consequences; log inputs, outputs, tool calls, and material state changes so that incidents can be reconstructed; and fold model failures into the incident-response process it already operates. Frameworks for all of this exist and are maturing, including the NIST AI Risk Management Framework and ISO/IEC 42001, and the striking thing about them is how unremarkable their contents are. They are recognizably the same disciplines that information security, privacy, compliance, and enterprise risk programs have practiced for decades, applied to a new class of software. That is the practical refutation of the mythology. A technology that yields to an inventory, a test plan, access controls, and an audit trail need not be treated as beyond governance, and the claim that it is can become an excuse for not doing the work.

Two kinds of organizations will emerge from this period, and the difference between them will not be who adopted the technology first. The organizations that accept the mythology will overpay for capabilities they have not tested, deploy systems they cannot explain into processes they cannot reconstruct, and discover their actual exposure in litigation, regulatory findings, or on the front page, at which point the vendor's language about emergent intelligence will offer them no shelter at all. Some of

them, chastened by the same mythology from the other direction, will freeze instead and forfeit real productivity to a fear of the miraculous that a competent assessment would have dissolved. The organizations that decline the mythology will do something far less dramatic. They will treat these systems as what they are: powerful and fallible software whose risks rise with what the system is permitted to remember, access, decide, and do. They will extract genuine value from them under controls they can defend to a regulator, an auditor, or a jury. The technology in the image that opened this paper has not been demonstrated, and fluent output is not evidence that a roadmap leads to it. What exists today is software, sold by vendors, assembled into systems, and deployed by people, and the organizations that remember this will own the outcomes they signed for, in both senses of the word.