philosophies

IMPACT FACTOR 0.7

CITESCORE 1.7

Article

The Backend as a Possible Functional Analogue of Consciousness: Redirecting Attention from the Language Model to the Orchestrating Layer

Pavel Straňák

Special Issue

Consciousness in the Age of Intelligent Systems: Philosophical Frameworks, Neural Theories, and Generative AI

Edited by

Dr. James Hutson and Dr. Luca M. Possati

MDPI

https://doi.org/10.3390/philosophies11030098

philosophies

MDPI

Article

The Backend as a Possible Functional Analogue of Consciousness: Redirecting Attention from the Language Model to the Orchestrating Layer

Pavel Straňák

Strategic AI Group, Czech Radio, Vinohradská 12, 12099 Prague, Czech Republic; pavel.stranak@rozhlas.cz

Abstract

Discussion of consciousness and artificial intelligence has hitherto focused on the question of whether a large language model (LLM) exhibits signs of consciousness or understanding. This paper proposes to redirect attention elsewhere: not to the model itself, but to the orchestrating layer that governs the model—the backend, understood here as the collection of mechanisms (context management, retrieval, evaluation, planning, and tool-use control) that structure the model's operation. We argue that the backend performs a function functionally analogous to the role of consciousness in the human brain: it stabilizes generative processes, directs attention, maintains context, and mitigates the entropic disintegration of thought. Consciousness fulfills this function through the phenomenal layer—qualia—which creates a persistent subjective "inner canvas", used here as a metaphor for a more general multimodal phenomenal space. The backend fulfills it only algorithmically, without phenomenal quality. We further show that computation is an informationally conservative process in the sense of Shannon's Data Processing Inequality (DPI), and therefore cannot increase Shannon information, even though it may yield novel or pragmatically useful recombinations of existing information. We conclude by proposing the hypothesis that consciousness constitutes a phenomenon orthogonal to computation—not an emergent property of complexity, but a qualitative leap into a different dimension. This hypothesis, which builds on the author's prior work in this Special Issue and in Symmetry, is presented as a conceptual contribution rather than a formal theory, and may have implications for how future artificial intelligence research conceptualizes the limits of computational architectures.

Keywords: consciousness; machine consciousness; philosophy of AI; large language models; backend orchestration; generative drift; phenomenal stability; orthogonal hypothesis; data processing inequality; philosophy of information; AI Theorem

Check for updates

Academic Editors: James Hutson and Luca M. Possati

Received: 30 April 2026
Revised: 5 June 2026
Accepted: 16 June 2026
Published: 17 June 2026

Copyright: © 2026 by the author. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction: We Are Looking at the Wrong Layer

When philosophy of mind and cognitive science turned to the question of artificial intelligence, they naturally directed their attention towards the language model. This is understandable: an LLM generates text that looks like thought, answers questions, translates, and argues. It therefore seems logical to ask whether this generator "understands" anything, whether it possesses intentional states, and whether it is or could be conscious. This paper is not a claim about the metaphysical nature of consciousness, but a conceptual exploration of structural analogies between computational architectures and biological cognition.

This paper proposes that the question may be posed at the wrong level of description. The LLM itself is a passive generator: it waits for input, produces the output of a single token, and without external control would halt after one token. Its outputs are, moreover, progressively subject to entropic decomposition. It is only the orchestrating layer—the deterministic backend—that provides any protection against this drift. A more detailed technical definition of the backend is provided in Section 2. This layer is often treated as a mere implementation detail. Philosophically, however, it is the crucial layer: it is the site where something functionally analogous to what we call consciousness in the brain takes place. In practice, system performance arises from the interaction between the LLM and the backend; modern techniques such as retrieval, tool use, and planning loops show that capability is a property of the whole architecture, not of the model alone.

Redirecting attention from the model to the backend is not merely a terminological adjustment. It opens a new way of thinking about the limits of computational intelligence and about the functional role of consciousness—a role that has not hitherto been explicitly formulated in this context. This paper builds on the author's earlier work, which analysed the ontological gulf between simulated and living intelligence [1].

2. Two Layers: Generative and Orchestrating

Modern AI systems have a clearly discernible two-layer architecture. At the lower layer lies the language model—a large neural network trained to predict the next token. This layer is associative, statistical, and passive. It produces a probability distribution over the vocabulary space; what it generates depends on the input context. It is a structurally stateless entity without persistent cognitive states: its cognitive "stability" is maintained only in the text itself, which repeatedly passes through it. The external orchestrating system of the backend must repeatedly send text to the model, append the generated token to its end, and transmit it again—the model itself, therefore, does not even run without the backend (Figure 1).

In this paper, the term backend refers to the full orchestrating layer that governs the operation of the LLM. Technically, it comprises system prompts, context-window management, retrieval and memory systems, tool-use pipelines, planning loops, evaluators, and multi-agent coordination mechanisms. Taken together, these components form a deterministic, explicitly engineered control layer that stabilizes and structures the inherently drifting generative process of the LLM.

Furthermore, autoregressive generation degrades over time: the model loses thematic focus, repeats itself, drifts from the task, or lapses into grammatically correct but semantically empty output. This phenomenon is referred to as generative drift.

At the upper layer lies the orchestrator—the backend. It decides when and how to invoke the model, what context to provide to it, how to interpret its output, and how to incorporate it into the next step. In contemporary systems, this backend comprises a heterogeneous set of mechanisms—system prompts, context-window management, retrieval pipelines, evaluators, planning loops such as ReAct [2], and multi-agent structures—that together provide continuity and task-directed control. In this sense, the backend is a fully deterministic, explicitly engineered control layer: a large, rule-like structure that governs when and how the passive generative function of the LLM is invoked. The backend is responsible for retaining prior context, filtering relevant information, and maintaining a focus the model itself cannot sustain. All current agentic architectures, from simple prompting to complex multi-step planning, are different realisations of the same function: the stabilisation of the generative process.

Figure 1. Two-layer structure characteristic of modern AI systems. The generative model at the lower layer (LLM—right) is a passive statistical engine: it produces one token at a time and has no mechanism for sustaining a coherent train of thought. The orchestrating backend at the upper layer (Backend—left) provides precisely those functions that the model lacks: continuity, context maintenance, relevance filtering, and goal-directed control. This division of labour is not an implementation detail but a structural feature of computational intelligence—and it has a direct analogue in biological cognition, where consciousness performs the stabilizing and integrative role that the backend approximates algorithmically.

This architecture has a direct structural counterpart in the brain. The neural circuits responsible for language production, associative thought, and memory retrieval are forms of the generative and associative layer. Consciousness, operating above them, fulfills the orchestrating function: it initiates, directs, filters, and integrates. This parallel is not merely metaphorical—it is functional: both systems require a higher-level integrative process to maintain coherence over time, even though the biological implementation includes a phenomenal dimension absent from artificial systems.

3. Consciousness as a Stabilising Mechanism: Qualia and the Inner Canvas

What precisely does consciousness do in the brain, and what distinguishes it from mere information processing? For the purposes of this paper, we focus on one specific function: consciousness creates what may be termed the "inner canvas"—a phenomenal space in which mental contents are present as experienced. Philosophically, this aspect of consciousness is described by the concept of qualia: the qualitative, subjective character of experience [3]. Chalmers's "hard problem of consciousness" [4] precisely articulates why this layer does not follow from a description of functional processes: there is no logical necessity for any quantity of computational processing to be accompanied by subjective experience.

From the standpoint of cognitive function, this inner canvas plays a crucial stabilising role. Baars [5], in his global workspace theory, likens consciousness to a theatrical stage on which selected mental contents are made available to the entire cognitive system. The experienced content—a visual scene, a verbal thought, an emotional tone—is present with a certain persistence and integrity. Consciousness appears to provide a stabilizing evaluative layer in which incoherent or impossible contents are rejected at a sub-symbolic level, not through textual analysis but through their appearance on the phenomenal canvas. Consciousness does not exhibit the rapid thematic drift characteristic of an un-orchestrated generative model; it is capable of sustaining focus, returning to an interrupted thought, and recognising incongruence between intention and output (Figure 2).

Figure 2. Phenomenal evaluation versus algorithmic evaluation. Left: The biological brain evaluates mental content on a phenomenal "inner canvas". When an impossible concept (e.g., a mouse with wheels) appears on this canvas, consciousness rejects it immediately as nonsensical at the sub-symbolic, perceptual level. The content is experienced as incoherent before it is analysed symbolically. Right: An AI system evaluates content only through symbolic and text-based operations. Both the LLM and the backend manipulate linguistic representations (e.g., simple conditional structures and text outputs such as 'conflict—nonsense, not real' shown in the illustration), but the system lacks any phenomenal canvas on which the content would appear as an integrated percept. Its rejection is therefore purely computational, not experiential.

This stabilising function has an important aspect: consciousness does not act as a textual filter, but as an evaluator at the sub-symbolic level. If a person entertains the thought of an "animal with wheels", consciousness immediately suppresses it as logical nonsense—not because it analyses text, but because the "animal with wheels" appears on the "inner canvas" as a nonsensical image. The intention is rejected by consciousness at a deeper level, more as a nonsensical percept than as nonsensical text. This property—integrated multimodal evaluation at the sub-symbolic level via the phenomenal inner canvas—has no fully adequate functional or phenomenal analogue in current AI systems [6].

The backend simulates this function algorithmically. It maintains the context window, filters relevant information, and injects reminders and corrections. It can thereby slow generative drift—but it cannot stop it entirely. Without phenomenal persistence, the backend must constantly re-inject externally the context that consciousness maintains naturally through experience. There is no "inner canvas" of the backend that could stabilise the direction of cognition; algorithmic stabilisation is a compensation for the absence of qualia—functionally useful, but never fully equivalent. In biological systems, this phenomenal layer operates in addition to non-conscious orchestrating processes, rather than as their simple extension.

The term "inner canvas" is intended as a concrete illustration of a more general phenomenal space. Although visual imagery provides an intuitive example, the phenomenal layer is not restricted to vision. Consciousness integrates multimodal experiential content—auditory, tactile, proprioceptive, affective, and conceptual—and remains fully present even in individuals who are blind from birth. The canvas metaphor therefore refers not to a literal visual field, but to the unified phenomenal arena in which mental contents are experienced.

In addition to providing this phenomenal evaluative space, consciousness is also mildly auto-excitatory: it can re-activate associative and orchestrating processes, sustaining an internal stream of thought rather than merely receiving it.

A simple everyday example illustrates this stabilising role of consciousness. When a person with wet feet thinks about the situation, consciousness does not drift into unrelated domains such as agriculture or hydrology. It keeps the mind anchored to the concrete experiential content—the sensation of moisture, the memory of stepping into a puddle, the intention to change shoes. Consciousness maintains the thematic thread because the relevant content is present on the phenomenal inner canvas. This phenomenal anchoring prevents associative explosion and keeps thought directed toward the actual problem at hand.

4. Generative Drift, Fixed Weights, and the Plastic Brain

Generative drift is not a technical defect correctable by a larger model. It is a characteristic tendency of autoregressive systems: each token influences the probability distribution for the subsequent token, and the cumulative effect of small deviations can lead to systematic divergence from the original intention.

The longer the AI generation, the more pronounced the drift. A typical illustration is provided by long ReAct traces [2], where an agent gradually loses direction and intent while preserving formal coherence.

It is notable that an LLM maintains grammatical coherence throughout generation, even during advanced semantic drift. The reason is structural: the model's weights are fixed during inference. The model cannot modify its parameters at runtime, and therefore the structural patterns from training—grammatical regularities, syntactic templates—are preserved as an inertial mechanism even when semantic content degenerates. Grammar is frozen in the weights; semantics drifts freely. The backend cannot fully prevent this by any sophisticated algorithmic method—what is missing is grounding in reality, intention, and a stabilising foundation: the inner canvas.

The behaviour of long autoregressive sequences can be understood through a physical analogy. A prompt injected into the model's vector space resembles a sharply bounded, intricately shaped cloud of smoke released into still air. With each iterative update—each microscopic "collision" in the generative process—the cloud expands, diffuses, and gradually loses its original structure. After sufficiently many steps, the initial shape dissolves into a uniform haze. Backend mechanisms act like improvised walls or directional

fans: they can slow the diffusion and momentarily redirect it, but they cannot restore the original form. In this sense, generative drift is a semantic diffusion process, and nothing within the computational architecture provides the phenomenal stabilisation that consciousness probably supplies in biological cognition.

This observation has a direct implication for the understanding of the brain. Living neural networks—brains—do not operate with frozen weights. Synaptic plasticity is their fundamental property: weights are continuously adjusted on the basis of experience and lived events. If an AI LLM were to modify its weights at runtime, it would lose its grammatical flywheel and the output would degenerate into pure noise. The plastic brain must solve this problem in a different way—and it appears to do so, at least in part, through consciousness: the phenomenal canvas provides a stabilising layer that in LLMs derives from fixed weights. Consciousness thus may function as a stabilising mechanism that complements plasticity. Without such a phenomenal anchor, plasticity would risk amplifying divergence rather than constraining it—yet biological cognition remains coherent, suggesting the presence of an additional stabilising dimension.

A thought experiment involving two LLM systems conducting a dialogue without a robust backend is instructive. Each drifts in its own direction; the grammar of both is preserved, but the shared topic disintegrates. A conversation that began with a goat ends with a cart—not because the models are unintelligent, but because there is nothing to anchor the thematic thread. And even if some new and meaningful theory were to emerge by chance in such an exchange, the models would not detect that they had reached their goal—as if they were playing ice hockey but there was no referee to recognise that a goal had been scored. The analogy of billions of cogwheels [1] applies equally at the level of dialogue: the number of cogwheels cannot generate a referee capable of distinguishing a goal from the puck passing through the goaltender.

5. Computation as an Informationally Conservative or Lossy Process

The existence of the backend resolves the stability problem to some extent, but it does not eliminate the deeper limit of computational intelligence. This limit is expressed by Shannon's Data Processing Inequality (DPI) [7]: no signal processing can increase the Shannon information contained in a signal. Information can be preserved or lost, but never created [8]. Computation is therefore informationally conservative or lossy—or more precisely, informationally closed—under deterministic transformations. In applying this principle to computational systems, we treat DPI as a conceptual constraint on how information flows through successive processing steps, rather than as a formal theorem about any specific architecture.

In the standard formulation, the DPI applies to abstract Markov chains $X \to Y \to Z$. Here X, Y, and Z represent successive informational states of a system, not physical variables or neural regions. In the context of an autoregressive model, X may correspond to the prompt, Y to the internal model state after processing it, and Z to the generated output. Each transformation can preserve or degrade Shannon information, but cannot increase it. This abstraction is sufficient to capture the behaviour of autoregressive inference, where each token update constitutes a step in such a chain.

For LLMs, this has a fundamental implication: a model can only recombine the information contained in its training data and input context. It cannot generate new Shannon information—that is, information that was not present in some form in its input. Sophisticated recombinations, interpolations, and extrapolations may look like creativity, but from an information-theoretic standpoint, they constitute a closed circuit. The contrary would be in conflict with the fundamental laws of information theory: sufficiently sophisticated signal processing could increase the information contained within it—and this would be

equivalent to a reduction in entropy without an external energy source, hence a violation of the second law of thermodynamics (Figure 3).

Figure 3. Computation as an informationally closed process. A mixer inside a fixed container symbolises the operation of an autoregressive model. The mixer continuously transforms the ingredients—data, words, and tokens—producing new combinations and patterns. Although the mixture changes its appearance, it never exceeds the boundaries of the container: no transformation can introduce information that was not already present in the initial ingredients. This illustrates Shannon's Data Processing Inequality (DPI): computation can preserve or degrade information, but cannot increase it in Shannon's sense. As the mixing continues, the structure of the mixture gradually decays, representing the entropic drift characteristic of generative models.

Although computation cannot increase information in the Shannon sense, it can reorganize existing information in ways that make it more semantically meaningful or pragmatically useful for an external observer. This distinction between informational quantity and pragmatic value is essential: a mixer can produce a better soup while simultaneously destroying the informational structure of the original ingredients. In the same way, an LLM may generate outputs that are more meaningful or useful, even though the underlying Shannon information has not increased. This does not violate the DPI; usefulness may increase, but Shannon information does not. For the user, however, the situation is reversed: if the user lacks the relevant knowledge (for example, how to cook a soup), the system's response increases the user's information, even if the model itself undergoes a small amount of internal drift or informational decay during the generation process. Crucially, this microscopic loss affects only the transient computation of a single answer; because the model's weights are frozen, the underlying structure of the system remains intact and no cumulative degradation occurs.

This idea may be illustrated by a limiting-case metaphor. Imagine a random text generator running for a sufficiently long time—in its output, the seed of a revolutionary theory would statistically appear. A physicist examining this output might recognise

it—the theory would not have occurred to her independently, but once she encounters it in random text, she can assess its value. Without an observer, the machine would continue running and the theory would be lost. An LLM is, compared to a random generator, more precise and structured by many orders of magnitude—this metaphor has the character of a limiting case, not a direct comparison. The principle, however, remains: if a model proposes a revolutionary theory, its revolutionary character cannot be competently evaluated within the same model. This requires something that is not an informationally closed circuit—something capable of assessing output from a position that is not part of the input data.

This aspect is related to what has been termed the absence of abductive reasoning [1]: AI does not know what it does not know, and cannot inquire, because it possesses no level of epistemic self-awareness. The abductive leap—the formation of a hypothesis across the boundary of available data—is not formally derivable from the information present in the input, and therefore represents precisely the type of new information that the DPI constrains within computational systems but which conscious cognition probably generates.

6. Epistemological Implications: What LLMs Have Revealed

The emergence of large language models has a paradoxical epistemological effect: they reveal what of thought is computationally reproducible—and thereby indirectly indicate what is not. LLMs have demonstrated that linguistic competence, apparent understanding, coherent argumentation, and creative recombination can arise statistically, without intentionality and without semantic grounding in the traditional sense. Much of what philosophy of language considered to be necessary conditions for linguistic performance has proven to be a historical artefact. LLMs thus paradoxically help to delineate the contours of what consciousness actually is, by showing what it is and what it is not [9].

At the same time, LLMs have revealed what probably cannot be computationally reproduced: enduring phenomenal stability and the capacity to generate genuinely new information in the Shannon sense. The backend, as an algorithm, can imitate the orchestrating function of consciousness [10], but cannot replace it—precisely because it lacks the inner canvas on which content is present as experienced. This absence is not a defect of implementation; it is structurally given. Related to this is the following claim derived from the Data Processing Inequality, which I refer to as the AI Theorem [7,11]. This formulation builds on my earlier work in Philosophies and Symmetry:

> *Any purely computational system whose reasoning trajectory is updated iteratively without external low-entropy input must eventually lose stable information and drift toward noise. In current LLMs this decay is confined to the transient inference process, as their parameters remain frozen and structurally unaffected.*

Frozen parameters are not an engineering convenience; they are an entropic necessity. By shifting the entropic burden from the model itself to the transient reasoning trajectory, the fixed weights of contemporary LLMs function as entropic crutches that prevent structural degradation. This claim is not directly grounded in empirical measurement, but is a conceptual consequence of the DPI and the entropic properties of autoregressive systems. We emphasize that the AI Theorem is introduced as a conceptual stability principle, rather than a formal mathematical theorem.

This perspective also clarifies how the present account differs from other approaches in information philosophy, such as Burgin's General Theory of Information [12], which extends the notion of information beyond Shannon's syntactic framework. The present argument remains within the Shannonian domain and concerns the limits of computational transformations under that constraint.

This changes the question we ought to be asking. Not: "Is the LLM conscious?"—but rather: "What must a system possess beyond the LLM and backend in order to achieve full

stability of thought and the capacity to produce deeply new information?" This question is both empirically and philosophically more productive, and it does not prejudge any of the competing theories of consciousness [6,10].

7. Consciousness as a Possible Orthogonal Phenomenon: The Hypothesis and Its Implications

From the foregoing analysis arises a fundamental question: is consciousness an emergent property of a sufficiently complex computational system, or is it a phenomenon of an ontologically different order? A hypothesis suggests itself, grounded not in metaphysics but in structural analogy: consciousness is orthogonal to computation. The term orthogonal is used here in its mathematical sense—as a relation between dimensions that cannot be reduced to one another—rather than as a metaphysical claim about substances.

The orthogonality hypothesis expresses the view that consciousness is not a computational phenomenon and cannot be reached by scaling or refining computational architectures alone. This claim is not metaphysical but structural: the phenomenal layer—the inner canvas on which experience appears—constitutes a functional dimension that is not generated by quantitative increases in computational complexity. In this sense, orthogonality denotes that no trajectory along the computational axis, however sophisticated, reaches the phenomenal axis.

This position does not deny that artificial consciousness may be possible, but it implies that its realization would require principles beyond computation as currently understood. Mechanisms such as persistent memory, vector databases, or multi-agent debate extend the computational axis horizontally; they do not introduce the phenomenal stabilizing dimension characteristic of conscious systems.

Placed among existing theories, the hypothesis complements rather than contradicts them. Whereas Integrated Information Theory grounds consciousness in intrinsic causal structure and Global Workspace Theory emphasizes access and broadcasting, the present account highlights the stabilizing role of phenomenal presence in preventing drift within a plastic cognitive system. Its contribution is to articulate why computational scaling alone may be insufficient and to motivate inquiry into architectures that incorporate a qualitatively different stabilizing layer [13,14].

Mathematics provides a useful analogy. The imaginary unit is not "more real numbers"; it represents a qualitative leap into a different dimension. No amount of real numbers, regardless of their magnitude or density, ever generates the imaginary axis. It must be introduced as a new axiom. The orthogonality hypothesis of consciousness makes an analogous claim: no amount of computational operations, regardless of their scale or sophistication, generates the phenomenal canvas. Consciousness is probably not the apex of a quantitative hierarchy of computational complexity; it lies on a different axis altogether (Figure 4) [15]. This view does not imply that artificial consciousness is impossible, only that it will not arise as a by-product of scaling present-day models. If consciousness is orthogonal to computation, its artificial instantiation would require a qualitatively different kind of architecture.

This view is philosophically consistent with Chalmers's formulation of the hard problem of consciousness [4], which highlights the explanatory gap between physical information processing and subjective experience. It also resonates with the ontological distinction between computational and living systems articulated in [1], where biological systems exhibit stability and coherence that cannot be fully captured by computational models alone. At the same time, this hypothesis does not deny the importance of quantitative factors such as neural scaling, hierarchical modularity, or network topology, which have been shown to correlate with cognitive capability in biological systems. Rather, it

suggests that such quantitative properties, while necessary for complex cognition, may not be sufficient to generate phenomenal experience.

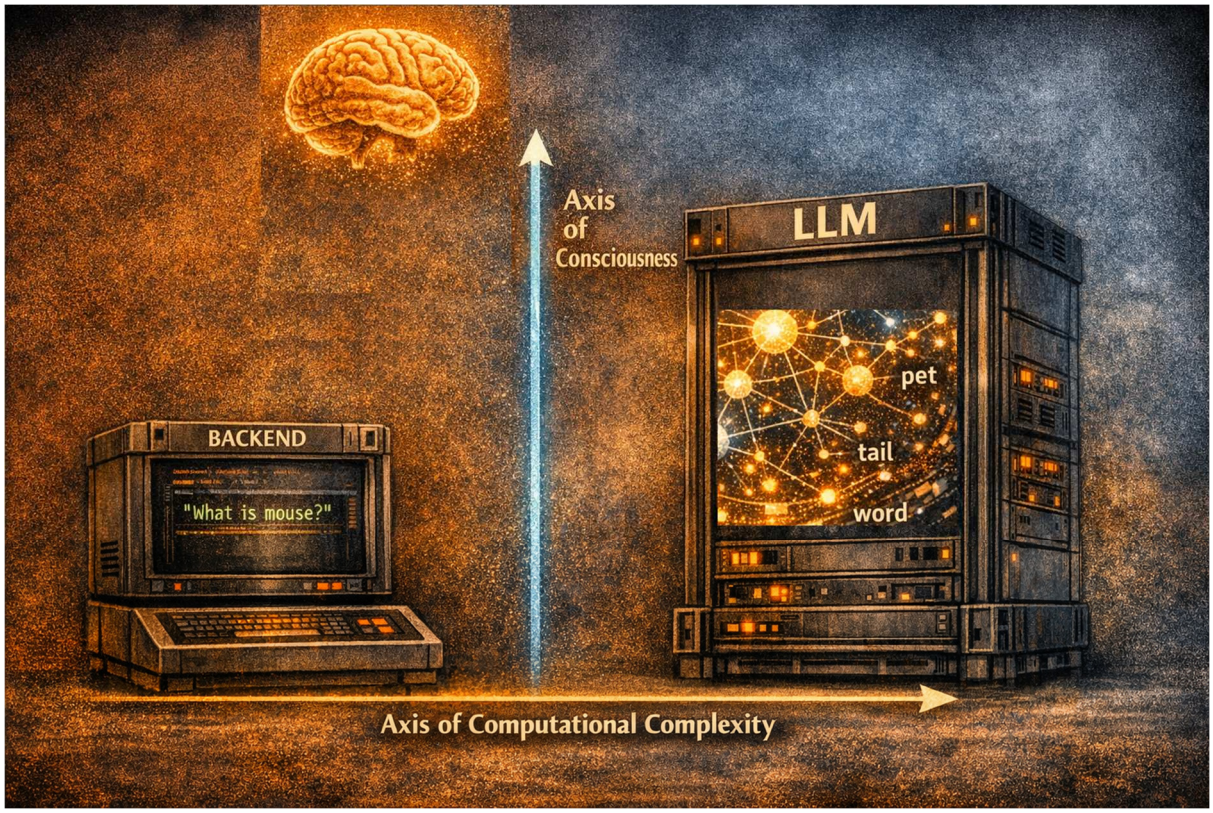

Figure 4. Possible orthogonality between computation and consciousness. The horizontal axis represents computational processes. At its lower end are contemporary artificial systems: the backend (left) and the LLM (right), both operating entirely within the computational dimension. The vertical axis represents the phenomenal dimension. The biological brain, with its conscious component, occupies a position above the computational plane, indicating that consciousness is not a quantitative extension of computation but a qualitatively orthogonal phenomenon. No amount of movement along the computational axis reaches the vertical axis—just as no accumulation of real numbers generates the imaginary axis. This illustrates the orthogonality hypothesis: consciousness lies on a different dimension than computation.

It is not appropriate to present this hypothesis as a demonstrable conclusion—current empirical and theoretical tools neither confirm nor refute it. Nevertheless, its implications for AI research are substantial. If consciousness is orthogonal to computation, then the pursuit of AGI through mere scaling of computational architectures may eventually encounter principled limits [16]. Attempting to reach consciousness by increasing computational complexity would be analogous to attempting to reach the imaginary axis by accumulating real numbers to infinity. On this view, the phenomenal dimension does not arise as a quantitative extension of the computational one.

This distinction becomes clearer when comparing biological cognition with contemporary AI systems. Although the brain must process linguistic and conceptual distinctions in a manner that is functionally similar to computation, its physical substrate is not computational in the digital sense. Neurons and synapses are not transistors, yet certain layers of neural processing behave in ways that resemble computational modules. This functional similarity should not be mistaken for ontological identity. In addition to its computational substrate, the brain contains a conscious component that stabilizes internal symbolic and associative processes (Figure 5). This stabilizing layer—the phenomenal canvas—anchors

thought, intention, and meaning in a way that prevents drift and preserves coherence over time. Contemporary LLMs, by contrast, lack such a stabilizing dimension. Their generative processes operate without an inner phenomenal anchor, and therefore require external orchestration to maintain coherence. The presence of consciousness in biological systems thus marks not merely a quantitative difference from current AI, but a qualitative one: a different kind of system, operating partly in a dimension that computation alone probably does not reach.

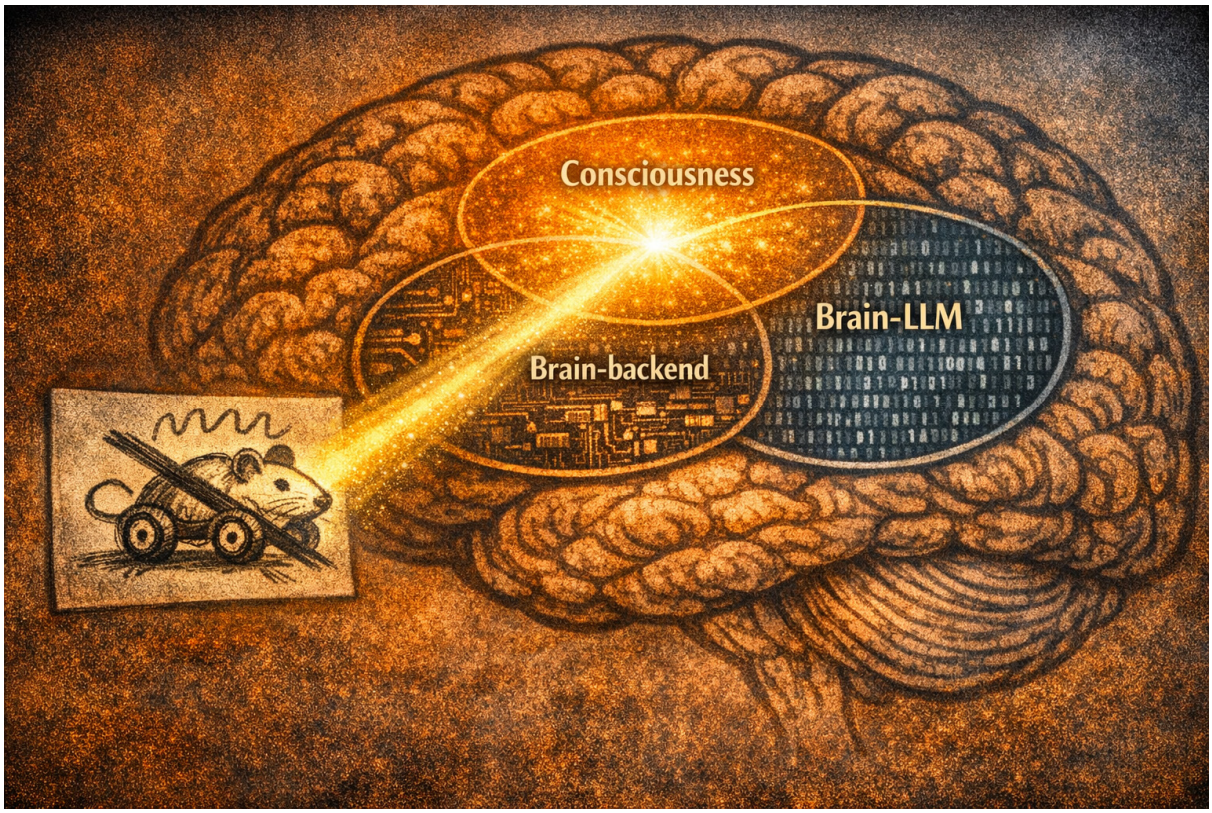

Figure 5. The brain contains three functional layers: a generative layer ("brain-LLM"), an orchestrating layer ("brain-backend"), and a phenomenal layer ("consciousness"). The phenomenal layer illuminates the inner canvas, where multimodal content appears as experienced. When an impossible concept (e.g., a mouse with wheels) is projected onto this canvas, consciousness rejects it immediately as incoherent, shown here by the crossed-out image. Beneath the phenomenal layer lie two layers that are computational in the functional sense, mirroring the roles of the backend and the LLM in artificial architectures. Unlike AI, however, the brain's conscious layer stabilizes and integrates the lower layers, preventing drift and maintaining coherent thought over time. The same three-layer structure applies to higher animals such as mammals: they possess a phenomenal layer and a basic generative–orchestrating substrate, although their "brain-LLM" consists only of limited symbolic and affective primitives rather than the rich linguistic and conceptual structures found in humans.

8. Conclusions

This paper has argued that contemporary AI systems should be understood as two-layer architectures in which the language model is only one component of a broader computational structure. The backend—comprising prompts, context management, retrieval, memory, planning, evaluation, and tool-use—provides the stability, direction, and coherence that the model itself cannot sustain. In biological cognition, analogous layers exist: a generative substrate of neural associations, non-conscious orchestrating processes, and, in addition, a phenomenal layer that supplies the stabilising inner canvas [17].

Phenomenal presence appears to play a functional role that current computational systems lack. Human consciousness is auto-excitatory: it reactivates and evaluates its own generative processes, producing a self-sustaining stream of thought. Artificial backends, when forced into self-activation, drift rather than converge. Biological plasticity further suggests that phenomenal experience may contribute a stabilising dimension: without such an anchor, continuously updating synaptic weights could amplify divergence. By contrast, LLMs maintain form through fixed weights but lack the phenomenal layer required to maintain content.

Shannon's Data Processing Inequality [7,8] shows that computation is informationally closed in the Shannon sense: it cannot create new information, only reorganize existing information. While such reorganization can yield semantically novel or pragmatically useful structures, computational systems cannot evaluate the significance of these structures within the same inferential loop. Human cognition, by contrast, can recognise and assess genuinely new ideas, suggesting that phenomenal experience may contribute a form of evaluative coherence not captured by computation alone [11,13–15]. A broader discussion of why contemporary AI architectures require an explicitly engineered orchestrating layer—and why this represents a return to engineering rather than an emergent computational solution—is provided in Appendix A.

These considerations motivate the orthogonality hypothesis: that consciousness constitutes a functional dimension not reached by quantitative increases in computational complexity. If this hypothesis is correct, then scaling existing architectures may encounter principled limits. This does not exclude the possibility of artificial consciousness, but it implies that its realisation would require principles beyond computation as currently understood. Research may therefore need to explore architectures with dynamic weights protected against entropic degeneration, internal feedback loops that simulate stabilising functions, or mechanisms not yet captured by computational models.

Biological evidence reinforces this structural distinction. Consciousness can vanish instantly under anaesthesia while autonomic processes continue; small mammals exhibit coherent behaviour only when conscious; and even minimal animals display affective responses that integrate sensory input with subjective evaluation [17]. These phenomena do not scale with brain size or computational capacity. Likewise, the behaviour of LLMs—despite their vast parameter counts—resembles a mechanistic transformation of stored human patterns, akin to a gramophone animating sound or a deterministic machine executing predefined operations. This architectural limitation may illustrate a principled boundary of computational systems [18,19]: without a phenomenal layer capable of evaluating and integrating content, computation alone cannot generate the directed, self-maintaining dynamics characteristic of conscious cognition.

The orthogonality hypothesis is therefore offered not as a metaphysical doctrine but as a conceptual tool for clarifying the structural limits of computational systems. If consciousness is indeed orthogonal to computation, then understanding its nature may not be the final step in artificial intelligence, but the first step toward a deeper theory of mind.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. This manuscript presents a theoretical framework and does not report empirical data.

Acknowledgments: The author thanks colleagues for discussions that shaped this work. Some passages of this manuscript, including figures, were prepared or refined with the assistance of a large language model (LLM, namely Microsoft Copilot version 2026). The author takes full responsibility for the content and conclusions presented herein.

Conflicts of Interest: The author declares no conflicts of interest.

Appendix A. The Return of Engineering

The emergence of large language models has often been interpreted as a decisive step toward machine intelligence. Unlike Turing's original conception of an intelligent machine as a vast, hand-crafted table of rules, LLMs are not engineered in detail but trained: their internal structure is not designed but "burned" into the weight matrix through exposure to large text corpora. This shift created the impression that intelligence had finally escaped the domain of explicit programming.

Yet the behaviour of an isolated LLM reveals a different picture. A transformer is essentially a static mapping from text to token probabilities—a vast, fixed structure that does not evolve during inference. It is a system that can generate language, but does not constitute a thinking process: it has no memory, no intrinsic dynamics, and no capacity to sustain a trajectory of reasoning. The early excitement surrounding LLMs arose from the illusion that fluent language implies fluent thought.

As these limitations became apparent, engineering returned to the centre of AI development. Modern systems rely increasingly on sophisticated backend architectures that perform retrieval, planning, evaluation, and context management. These components are fully engineered, deterministic, and explicitly designed—precisely the kind of structures that the emergence of LLMs was thought to render obsolete. In practice, the quality of the backend now determines the apparent intelligence of the system more strongly than the size of the model itself.

This development raises a deeper question: have we moved closer to artificial thought, or merely achieved a powerful automation of language? The evidence suggests the latter. LLMs provide an unprecedented linguistic substrate, but the coherence, direction, and stability of their outputs arise from engineered scaffolding rather than intrinsic cognitive dynamics. What appears as reasoning is often a composite of statistical generation and algorithmic orchestration, not a unified process of thought.

The trajectory of AI thus oscillates between emergence and engineering. LLMs demonstrate that vast structures can be learned rather than designed, but the systems that appear intelligent still rely on carefully constructed mechanisms external to the model. Whether this hybrid architecture can ever converge toward genuine cognition—or whether consciousness requires a fundamentally different kind of process—remains an open question.

At present, only conscious human agents—endowed with phenomenal awareness and goal-directed evaluation—can generate genuinely new principles that advance AI architectures. Contemporary AI systems can assist this process, sometimes dramatically, but they do so in the manner of powerful tools: they amplify human labour without possessing any understanding of where the work should be directed.

As backend architectures increasingly integrate real-time retrieval and tool use, the need to encode vast internet-scale data corpora directly into the model's weights diminishes. Future systems may rely less on massive monolithic models and more on modular architectures in which smaller LLMs collaborate with engineered components that perform search, evaluation, and planning. This development highlights the rise of an 'artificial awareness' within backends—a form of system-level competence that is critically important for overall performance. Yet this artificial awareness remains entirely algorithmic and therefore fully engineered. The circle thus closes: unless the principles of genuine consciousness

are discovered and instantiated in machines, engineering will remain the dominant force shaping AI systems.

References

1. Straňák, P. What Artificial Intelligence May Be Missing—And Why It Is Unlikely to Attain It Under Current Paradigms. *Philosophies* **2026**, *11*, 20. [CrossRef]
2. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the 11th International Conference on Learning Representations, ICLR, Kigali, Rwanda, 1–5 May 2023.
3. Nagel, T. What is it like to be a bat? *Philos. Rev.* **1974**, *83*, 435–450. [CrossRef]
4. Chalmers, D. Facing up to the problem of consciousness. *J. Conscious. Stud.* **1995**, *2*, 200–219.
5. Baars, B.J. *In the Theater of Consciousness*; Oxford University Press: Oxford, UK, 1997.
6. Seth, A.K.; Bayne, T. Theories of consciousness. *Nat. Rev. Neurosci.* **2022**, *23*, 307–321. [CrossRef] [PubMed]
7. Shannon, C.E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
8. Cover, T.M.; Thomas, J.A. *Elements of Information Theory*, 2nd ed.; Wiley: Hoboken, NJ, USA, 2006.
9. Dehaene, S. *Consciousness and the Brain*; Viking Press: New York, NY, USA, 2014.
10. Tononi, G. An information integration theory of consciousness. *BMC Neurosci.* **2004**, *5*, 42. [CrossRef] [PubMed]
11. Straňák, P. Possible Entropic Limits of Iterative Computation in Generative AI. *Symmetry* **2026**, *18*, 764. [CrossRef]
12. Burgin, M. *Theory of Information: Fundamentality, Diversity, and Unification*; World Scientific: Singapore, 2010.
13. Butlin, P.; Long, R.; Elmoznino, E.; Bengio, Y.; Birch, J.; Constant, A.; Deane, G.; Fleming, S.M.; Ji, X.; Kanai, R.; et al. Consciousness in Artificial Intelligence. *arXiv* **2023**, arXiv:2308.08708.
14. Schwitzgebel, E. AI and Consciousness. *arXiv* **2025**, arXiv:2510.09858.
15. Levine, J. Materialism and qualia: The explanatory gap. *Pac. Philos. Q.* **1983**, *64*, 354–361. [CrossRef]
16. Ardila, A.; Bernal, B.; Rosselli, M. Executive functions brain system: An activation likelihood estimation meta-analytic study. *Arch. Clin. Neuropsychol.* **2018**, *33*, 379–405. [PubMed]
17. Herculano-Houzel, S. Numbers of neurons as biological correlates of cognitive capability. *Curr. Opin. Behav. Sci.* **2017**, *16*, 1–7. [CrossRef]
18. Salatiello, A. Modularity is the bedrock of natural and artificial intelligence. *arXiv* **2026**, arXiv:2602.18960.
19. Bassett, D.S.; Greenfield, D.L.; Meyer-Lindenberg, A.; Weinberger, D.R.; Moore, S.W.; Bullmore, E.T. Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS Comput. Biol.* **2010**, *6*, e1000748. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.