EPHAPTIC INDISPENSABILITY:
A NEURAL KILL SWITCH FOR AI AGENTS

Ismaila "izzo" Wane

June 15, 2026 (Rev. 0.3.1)

ABSTRACT

This paper establishes that an architecturally enforced neural kill switch is possible by implementing a native governance credential that is structurally necessary for authorized inference, such that its removal produces provably unauthorized outputs by a lower-bounded margin. We call this property the Ephaptic Indispensability Guarantee, and we implement it via a biologically-inspired architecture that embeds governance directly into the activation-space forward pass. We establish the guarantee along three independent axes. Biologically, we ground the claim in neuroscientific evidence that ephaptic fields, the brain's electric fields that directly influence individual neurons, are not auxiliary signals but load-bearing control parameters; their disruption in biological neural networks produces measurable, clinically significant cognitive breakdown. Mathematically, we formally state and prove the theorem under an explicit construction, establishing a lower-bounded separation between authorized and unauthorized inference, and deducing that parameter-space approaches such as QLoRA cannot satisfy this bound in any deployment mode. Empirically, we validate the theorem on Qwen3.5, a state-of-the-art open-weight language model measuring distributional divergence, task accuracy, perplexity, and hidden-state trajectory distance under various unauthorized conditions. We demonstrate model sovereignty by binding the modeled ephaptic fields to hardware-attested credentials and enforce secure inference governance over any model, regardless of origin. Together, these results establish ephaptic coupling as the first mathematically proven, empirically validated architectural kill switch for AI agents, one where governance is woven into the model's neural dynamics, not bolted on externally.

1. INTRODUCTION

1.1 Motivation

Organizations are presently expecting verifiable controls for safety, privacy, provenance, and cost before deploying AI agentic systems and their associated models at scale. This expectation is no longer just aspirational as regulators in various countries and regions around the world are actively codifying it into laws.

In the United States, the Department of Commerce's Center for AI Standards and Innovation (CAISI) [26] now conducts pre-release safety evaluations of frontier AI models, having already blocked the public release of state-of-the-art models it deemed unsafe, operating against the NIST AI Risk Management Framework [27] as a technical baseline. Further, a White House executive order issued on June 2, 2026 [38] establishes a voluntary pre-release assessment framework for designated frontier AI models. The United Kingdom's AI Safety Institute [28] runs a pre-deployment testing regime in coordination with leading frontier developers. Canada's national AI strategy, launched on June 4, 2026 [39], similarly expands the Canadian AI Safety Institute's capabilities to conduct transparent evaluations of AI models, commits to deploying trusted AI agents at national scale, and modernizes legislative frameworks to prioritize trust, safety, and protections against harmful AI-generated content. China's Cyberspace Administration has issued algorithmic transparency requirements and binding labeling rules for AI-generated synthetic content [29]. The European Union's AI Act [30] establishes binding, risk-tiered obligations for AI providers and deployers, including conformity assessment, transparency, and post-market monitoring for high-risk systems, with general-purpose AI model obligations already applicable since August 2025 and high-risk AI rules phasing in across August 2026 and August 2027. The stakes are now even higher with concerns extending beyond industry and the state. In May 2026, Pope Leo XIV devoted his first encyclical, Magnifica Humanitas [33], to AI with a dedicated section on "Responsibility, transparency and the governance of AI", concluding that *"merely regulating it is insufficient; it must be disarmed"*, by which he meant freeing it from monopolistic control and preventing it from dominating humanity, framing AI governance as an obligation for all of humanity. When a moral voice addressing billions of people frames AI governance as a question for humanity, not merely for firms or regulators, the requirement shifts from controls that are declared to controls that are structurally enforceable.

We refer to the underlying technical challenge of providing controls that survive the model leaving the guarded surface as the kill switch problem. The kill switch can be further defined as a mechanism that can halt, constrain, or redirect agent behavior under adversarial conditions at the discretion of the deploying organization or its intended operator.

The premise of the AI kill switch is that if a system can be reliably stopped or overridden, its risks are bounded. Yet no existing approach provides this guarantee at the neural layer. API guardrails, output classifiers, policy wrappers, and access-control mechanisms are applied around the model rather than within the computation itself. Any adversary, or sufficiently capable agent, that can operate the model outside the protected surface can bypass these controls entirely. The off button exists, but it remains external to the computation it is intended to govern, leaving the kill-switch problem unresolved. The kill switch is therefore emerging as one of the defining technical and challenges of the AI era. In a nutshell, it is arguably a generational problem, a "challenge of our time" [33].

That said, the difficulty is much deeper than simply adding an "off button". This is why many current AI kill switch discussions are still superficial. Indeed, most proposals assume centralized infrastructure control, but the real problem emerges once intelligence becomes decentralized, persistent, and independently reproducible. The "off button" is very hard to implement on a neural network emulating a human brain and that is precisely why the problem is fundamentally different from traditional software control. A sufficiently advanced artificial neural network is not just executing deterministic instructions line-by-line like classical software. It is learning distributed representations, generalizing, adapting behavior, and producing emergent outputs from billions or trillions of parameters. In that sense, the "knowledge" or "capability" is diffused across the network, somewhat analogous to biological cognition. There is no single symbolic rule to remove which creates several difficulties. Indeed, one cannot simply "delete" one function to remove a capability. Behaviors may re-emerge after fine-tuning. Safety constraints can degrade under distribution shift. Distilled or quantized descendants may preserve dangerous capabilities. The same weights can behave differently depending on prompting, tools, memory, or agentic scaffolding. This is closer to governing a cognitive system than managing conventional software binaries. A useful biological analogy is that one can physically stop a biological brain, but cannot easily remove a specific thought pattern, capability, or behavioral tendency once encoded into neural structure. Having such capability would effectively equate to some form of "mind control".

Similarly, with frontier AI, the model weights themselves embody latent capability, and once replicated, the "mind" effectively persists beyond a single infrastructure instance. That is why the AI kill switch problem becomes a cryptographic challenge, a hardware trust challenge, a distributed systems challenge, and ultimately a governance challenge. It cannot be merely an application-layer feature. In fact, the more artificial neural networks (ANNs) approximate cognitive architectures found in biological systems, the less compatible they become with traditional centralized shutdown assumptions inherited from classical software engineering and security approaches.

This generational technical challenge warrants inquiring if synaptic weights alone, or broadly neural dynamics, can govern trustworthy AI agents. This paper proves that such a kill switch at the neural level is not only possible, but can be made architecturally necessary, woven into the model's forward pass such that bypassing it is not a matter of circumventing an external mechanism, but of making the model produce provably unauthorized outputs. We call this property the Ephaptic Indispensability Guarantee. The native governance credential, emulating ephaptic fields and bound to a hardware-attested keyed state, is not just a switch attached to the model. It is essentially a structural component of what the model computes. Removing it does not leave the model intact and ungoverned; it leaves the model producing outputs separated from authorized inference by a lower-bounded margin $\delta > 0$ on a non-zero measure of inputs. This distinction matters greatly. Indeed, a kill switch that can be bypassed is a liability. It can provide the appearance of control without the substance. A kill switch that is architecturally enforced is a guarantee. This paper establishes the first such guarantee.

As previously noted, existing approaches are typically external to the model. A sufficiently capable or adversarially controlled agent can circumvent external controls. Recent work has explored representation-level defenses. For instance, Rosati et al. [10] demonstrate that harmful capabilities can be noised out of model representations to resist malicious fine-tuning, operating across layers in representation space. However, Rosati et al. [15] subsequently prove that convergence-rate control defenses, which attempt to make fine-tuning attacks computationally expensive, have fundamental limits as adversaries can restore fast convergence through proportional model expansion. These results reinforce therefore a deeper structural point highlighting that defenses that operate within the weight or parameter space, however sophisticated, face inherent ceilings because the base model weight matrix W remains a complete, functional computation. This leads us to initially pose the following important question: can governance be made architecturally necessary (i.e., mathematically inseparable from inference)?

After privately pursuing the concept for nearly a decade, Wane [12] introduced Ephaptically Coupled Artificial Neural Networks (EC-ANNs) in June 2025, demonstrating ephaptic coupling as a novel neuromodulation mechanism, operating on activation space to improve performance and stability of traditional artificial neural networks while outlining future security work. Indeed, the original EC-ANN paper specifically notes that the ephaptic coupling mechanism, when implemented through an adaptive intelligent system, raises significant ethical and security concerns related to emergent behavior in AI systems. It was articulated that artificial ephaptic fields that modulate and adapt in real time after each inference, or over the course of deployment, could result in AI agents that evolve beyond their original specifications. Such adaptability introduces the risk of rogue or compromised AI agents, either through unintentional behavior or deliberate misuse by bad actors. Wane approached this possibility with caution and pointed to some future work developing cryptography-based safeguard mechanisms that are tamper-resistant and resilient to future threats. This follow-up paper closes that security gap, by essentially treating ephaptic coupling as an inherent neural control plane with ephaptic fields modeled as an Ephaptic Coupling Matrix (ECM) denoted as Λ, which operates on activation space and governs how activations propagate.

Ephaptic coupling enables three governance properties, as summarized in Table 1 which are not available from any external guardrail or policy wrapper as they require the governance credential to be woven into the forward pass itself:

Property	Definition
Identity and provenance-bound modulation	Λ is encrypted and cryptographically bound to the deploying agent's hardware credential (K) through a personalization process; it cannot be decrypted or applied outside the attested host (cloud or edge device) .
Governance control	Enabling, disabling, and revoking inference is effectuated by invalidating K, not by blocking an API endpoint, which can be circumvented. This provides on-demand remote control.
Tamper detection and resistance	Because Λ is structurally necessary for authorized inference, any tampering with the model or its execution environment produces outputs that provably diverge from authorized behavior

Table 1: Governance Properties

A related, increasingly urgent motivation is also the model sovereignty problem. Indeed, organizations are increasingly deploying high-capability models originating from third parties, including models developed by laboratories in foreign jurisdictions with different regulatory environments. Models such as Qwen 3.5 (Alibaba)[14], Llama (Meta), Mistral, and others offer state-of-the-art performance and are often open-weight, making them attractive for deployment in sensitive or regulated environments. Yet organizations deploying these models face a structural governance problem as a model's weights are controlled by the model provider, not the deployer. External wrappers (e.g., API guardrails, safety classifiers, prompt filters) are not part of the model and can be bypassed, replaced, or stripped. The deploying organization has no architectural guarantee that its governance policy is enforced at inference.

Ephaptic indispensability construction directly addresses this gap. By injecting Λ into the model's forward pass and binding the ECM Λ to the deploying organization's root-of-trust credential, the deployer establishes cryptographic governance that is architecturally necessary. It cannot be removed without producing outputs that provably diverge from authorized inference by a lower-bounded margin on a non-zero measure of inputs. This applies to any model, regardless of origin. In other words, the deploying organization becomes the cryptographic governor of the model's behavior, independently of the model provider. This principle, referred to as model sovereignty, is essentially the right, obligation and the technical capability of the deploying organization to hold architectural control over AI systems they operate, regardless of where those models were built. The additional question of whether weight matrix W alone can reproduce authorized behavior without such ECM Λ remains then open. This paper closes that question by proving the Ephaptic Indispensability Guarantee.

1.2 Contributions

This paper makes the following key contributions:

- We ground the theorem in neuroscientific evidence connecting ephaptic field necessity in biological neural networks [7, 9, 11] to the architectural indispensability of the ECM Λ in artificial neural networks, establishing a biological precedent for the mathematical claim.

- We formally state the Ephaptic Indispensability Guarantee, a theorem establishing that there exists an EC-ANN construction under which the synaptic weights W alone cannot reproduce authorized inference behavior, with a provable lower-bounded separation $\delta > 0$ on a non-zero measure of inputs.

- We provide a constructive proof under the Partial-Path Completion and Indispensability Training Objective construction (Families C + D). Every assumption in the proof is explicitly stated, independently verifiable, and either enforced by construction or derived from task complexity.

- We prove that LoRA-based adapters cannot satisfy the indispensability lower bound in any deployment mode (i.e., merged, standalone, or keyed) establishing that parameter-space mechanisms are structurally incapable of serving as governance primitives.

- We present the first empirical validation on a state-of-the-art open-weight model (Qwen 3.5) across language tasks, closing the scale gap from prior EC-ANN results on GPT-2 small. Cross-domain coverage via reinforcement learning and vision extends the generality of the construction.

- We introduce the model sovereignty framing whereby EC-ANN governance is enforced as a deployer-held cryptographic property that applies to any model regardless of origin, enabling organizations to hold architectural control over third-party models deployed in their infrastructure.

1.3 Relationship to Prior Work

Although the addressed threat model is different, there are various state-of-the-art AI defense systems worth mentioning with regard to our wok. Representation Noising (RepNoise) from Rosati et al. [10] perturbs activations during alignment so that harmful capabilities become difficult to recover via fine-tuning, while leaving the underlying weight matrix W as a complete forward path. Tampering Attack Resistance (TAR) from Tamirisa et al. [31] were the first defense shown to withstand a meaningful number of open-weight fine-tuning attacks while preserving capability; TAR hardens W and the refusal/unlearning safeguards baked into it against tampering. Vaccine from Huang et al. [32] introduces perturbation-aware alignment, producing hidden embeddings that resist harmful-prompt-induced drift during downstream user fine-tuning. All three represent state-of-the-art for their respective threat models, and all three leave the weight matrix W as a complete forward path. They make recovering the unsafe model harder, but they do not make W structurally incomplete. The architectural-impossibility argument developed in this paper applies uniformly to them (i.e., an adversary who obtains W has, in every case, a complete forward path). The Ephaptic Indispensability Guarantee occupies a different point in the design space. It is not a stronger weight-space defense, but a structural relocation of the governance primitive from weight space (W) to activation space (Λ), where its absence is detectable as a lower-bounded margin on outputs, rather than indirectly through how much harmful capability an adversary can recover from W.

The original EC-ANN paper [12] established cross-domain performance gains but left indispensability implicitly for future work. It did, however, explicitly recommend developing tamper-resistant cryptographic safeguards in parallel to address ethical and security concerns. Granted patent [13] and pending patent applications already disclose the EC-ANN architecture alongside an accompanying virtual modulation device, that discovers the optimal hyperparameters values via a Bayesian search, and a public key infrastructure (PKI)-bound deployment, that provides hardware-backed cryptographic security. A claim specifically covers the use of a cryptographic key stored in a secure element as input to the EC-ANN activation function which is the patent-level articulation of what this paper formalizes. Another claim explicitly covers a “kill switch feature” implemented by preventing intra-layer communication upon authentication failure. This paper provides the scientific foundation that explains why preventing ephaptic coupling produces provably unauthorized outputs, giving the patent claims their theoretical grounding. While the patent family protects a wide variety of systems and methods for implementing the ephaptic coupling mechanism into the field of AI, it is worth noting that both original and herein follow-up papers aim to solely provide the theoretical framework and empirical evidence to support the novel mechanism.

2. BIOLOGICAL BACKGROUND

This section establishes that ephaptic fields are not auxiliary signals in biological neural networks. They are effectively load-bearing control parameters whose disruption produces measurable cognitive breakdown. This biological precedent motivates the mathematical construction in Section 3.

2.1 From Synapse to Ephapse

The term "ephapse" was coined by Arvanitaki [1] in 1942 to describe sites of functional interaction between adjacent nerve fibers mediated by local electric fields rather than synaptic transmission. It is only in 2011 that Anastassiou et al. [5] confirmed ephaptic coupling in mammalian cortex, demonstrating that extracellular electric fields generated by neural activity can modulate the firing of neighboring neurons. More recently, Cunha et al. [7] and Pinotsis et al. [9] have shown that ephaptic coupling plays a significant role in neural synchronization and network complexity. Since the invention of the perceptron in 1957 by Rosenblatt and later formalized in 1962 [2], standard artificial neural networks primarily model the synaptic component of neural communication. Each neuron typically computes a weighted sum of inputs from the prior layer (a row-wise operation on the weight matrix W). The ephaptic component, within-layer, column-wise modulation mediated by local fields, is entirely absent from conventional artificial neural network architectures. Yet in biological neural tissue, ephaptic interactions may rival or exceed synaptic connections in dense cortical regions such as the hippocampus, neocortex, and cerebellum.

2.2 The Cytoelectric Coupling Hypothesis

Pinotsis et al. [9] proposed the cytoelectric coupling hypothesis which suggests that efficient information processing in the brain requires stable mesoscale electric fields that organize neural activity at a scale between individual neurons and large-scale brain regions. Under this hypothesis, electric fields act as control parameters in the sense of Haken's synergetics [3], slowly-varying fields that enslave fast-relaxing neural activity through circular causality. The field does not merely reflect neural spiking; it organizes it. Perturbing the field changes both the field and the neural activity simultaneously, because the two are coupled in a bidirectional feedback loop. This is the critical structural insight for the indispensability argument. The field is not downstream of computation, but it is a critical component of what computation means in the biological system.

2.3 Evidence from Depression

It has been scientifically demonstrated that ephaptic disruption breaks neural computation. Indeed, Pinotsis et al. [11] published a fascinating paper titled "*Ephaptic coupling and power fluctuations in depression*", where they study data from an experiment with a small patient cohort with treatment-resistant depression. The data included recordings from subcallosal cingulate cortex over four weeks following intraoperative deep brain stimulation (DBS). We highlight in Figure 1 the transition from stable attractors to itinerant states by illustrating our synthesis of the dynamical properties identified by Pinotsis et al.,

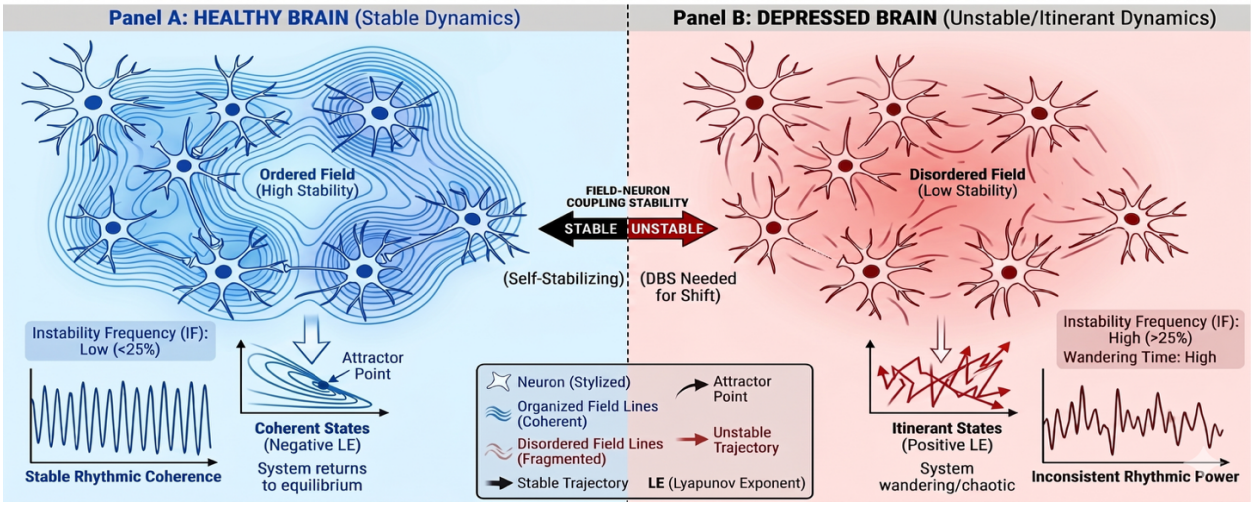

Figure 1: Ephaptic Coupling in Healthy vs. Depressed Brains

Furthermore, three additional key findings emerge from the work of Pinotsis et al.:

- Field stability: Electric field Lyapunov exponents consistently lower than neural activity exponents so fields are more stable, confirming slaving principle prediction (i.e., Pinotsis's "conductor-orchestra" analogy)
- Disrupted fields correlate with severity: Two new measures, relative wandering time and instability frequency, track depression scores in left hemisphere:
    - Mild depression: negative relative wandering time, instability frequency <20%
    - Severe depression: positive relative wandering time, instability frequency >20%
- Power fluctuations are epiphenomenon of itinerancy: Lyapunov exponents correlate with oscillatory power. Brain power fluctuations reflect itinerant dynamics with the brain wandering between attractors.

These findings from Pinotsis et al. allow us to infer a biological indispensability argument of ephaptic coupling. In other words, removing or disrupting ephaptic field coordination does not leave synaptic computation intact. Instead, it breaks the control parameter that organizes neural activity. Simply put, spikes without the field cannot maintain coherent computation. The biological analogue maps directly with the core idea behind EC-ANNs. Indeed, the ECM Λ plays the role of the ephaptic fields while W plays the role of synaptic transmission. If removing the biological ephaptic fields breaks computation in vivo, the question is whether a constructed Λ can be made indispensable in the same structural sense in silico. This motivates the formal theorem in later sections.

It is also worth noting Rosenblatt [2] has declared in his quest to develop brain models, that it will be a general strategy to start out with minimally constrained networks, and examine the consequences of introducing particular types of constraints, one at a time. Rosenblatt noted that brain phenomena require "built-in" control mechanisms, of a rather intricate sort, noting that these built-in mechanisms were not known in any detail at the time. Evidence available at the time from elementary functions prompted Rosenblatt to make conjectures of the types of "computational mechanisms" that are likely to exist throughout the central nervous system. After a thorough review of Rosenblatt's work, we found no evidence that he was aware of Arvanitaki's work. This could explain why ephaptic coupling was not incorporated in Rosenblatt's foundational brain model or in *Principles of Neurodynamics*, both of which focused solely on synaptic communication. This may also explain why ephaptic coupling has been historically overlooked by the AI community.

3. MATHEMATICAL BACKGROUND

To appreciate ephaptic indispensability from a mathematical perspective, it is important to address first why parameter-space approaches cannot achieve indispensability.

3.1 Limitations of Parameter-Space Alternatives

Low-Rank Adaptation (LoRA; Hu et al. [6]) is the standard approach for parameter-efficient fine-tuning. Rather than updating the full weight matrix W, LoRA trains two low-rank factor matrices $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$ with rank $r \ll \min(d, k)$, such that the adapted output is:

$$y = (W + s \cdot BA)\, x$$

Quantized LoRA (QLoRA), as implied, extends LoRA by quantizing the frozen base weights W to 4-bit precision while keeping A and B in full precision, making large-model fine-tuning feasible on consumer hardware. It is worth noting that LoRA adapters operate purely in weight space as they only modify or augment the synaptic term W, which governs how information flows row-wise across layers.

We establish that LoRA-based fine-tuning operates in the wrong space to serve as a governance primitive, and why activation-space modulation is the correct locus for architectural indispensability. It is important to note that we do not argue that EC-ANNs are superior to LoRA-based mechanisms, which are highly effective, widely validated tools for parameter-efficient fine-tuning. LoRA's achievements surpass today EC-ANNs in their current form as they have not yet been fully validated on at scale. However, they serve a fundamentally different purpose than the one examined here. The argument is narrower and more precise, focusing on the specific property of architectural governance making a credential structurally necessary for authorized inference. LoRA adapters cannot serve as the primitive, regardless of how they are keyed or encrypted. This is a structural claim, not a performance claim. EC-ANNs with the

indispensability construction achieve something LoRA cannot. The two mechanisms are complementary, and the honest comparison is on the governance axis alone. The fundamental limitation is that optional enhancements are not governance primitives. The comparison provided in Table 2 reveals the core limitation of LoRA adapters.

Mode	What happens to A, B	Base model without A, B
Merged	BA is folded into W permanently	Identical to fine-tuned model, adapters no longer exist as separate artifacts
Standalone	A and B applied at inference as $W + s \cdot BA$	Fully functional, degrades capability, but inference proceeds normally

Table 2: Deployment Modes of LoRA Matrices

The absence of A, B does not break the model. W is a complete, functional neural network without the adapters. Removing them degrades quality but does not scramble computation. The base model simply runs as it was pretrained. This is why LoRA matrices cannot serve as cryptographic governance primitives, regardless of how they are keyed or encrypted. They are optional enhancements in weight space, and not architectural dependencies in the forward pass. An adversary who obtains W has a working model. The governance layer is detachable by construction. This limitation is not incidental, but structural. LoRA operates exclusively on the synaptic term, via row operations: each hidden neuron's output is a dot product of its weight row with the input. Modifying this row-wise quantity, even with a cryptographically bound delta, still leaves W as an architecturally complete forward path. The adapter contributes to what the model knows; it does not govern whether the model can compute. This structural ceiling extends beyond LoRA to any defense mechanism that operates within parameter space. Rosati et al. [10] demonstrate that representation noising can remove harmful capabilities from model activations, but Rosati et al. [15] prove that such convergence-rate control approaches have fundamental scaling limits: an adversary can always overcome the defense by proportionally expanding the model. The root cause is the same as W remains a complete forward path. The defense adds friction and does not add architectural dependency. The dimensional distinction (i.e., row space vs column space) is key to understanding this limitation. Indeed, standard ANNs (and by extension LoRA) operate entirely in the row space of the weight matrix W:

$$S_i = \sum_j W_{ij}\ x_j + b_i \qquad \text{(synaptic, row-wise, cross-layer)}$$

Where:

- S_i is the synaptic term for neuron i
- W_{ij} is the synaptic weight from neuron j to neuron i
- x_j is the post-synaptic activation of neuron j from the previous layer
- b_i is the bias term for neuron i

Each row $W_{ij:}$ defines one neuron's receptive field from the prior layer. LoRA's $\Delta W = BA$ introduces a low-rank perturbation in this same row space. It is an efficient approach but fundamentally homogeneous with W.

3.2 Ephaptic Coupling as Activation-Space Alternative

As modeled by Wane [12] in the context of artificial neural networks, ephaptic coupling introduces a second, orthogonal axis for performing column-wise modulation within the same layer:

$$E_i = \varepsilon \sum_k \Lambda_{ik}\ \Phi(x_k) \qquad \text{(ephaptic, column-wise, intra-layer)}$$

Where:

- E_i is the ephaptic coupling term acting on neuron i
- ε is the ephaptic factor scalar controlling the strength of ephaptic modulation
- Λ_{ik} is the ephaptic coupling coefficient from neuron k to neuron i
- Φ is the nonlinear transformation (e.g., Identity, SiLU, ReLU, tanh) of x_k
- x_k is the post-synaptic, pre-ephaptic activation of neuron k within the same layer as neuron i

It is worth noting that the subscript indices (i, j, k) used throughout this paper encode three orthogonal axes of neural computation. The synaptic term S_i uses indices i (output neuron) and j (input neuron from the previous layer) to span inter-layer signal propagation. The ephaptic term E_i introduces a third index k, denoting a peer neuron within the same layer. It indexes intra-layer coupling (i.e., columns of Λ), not inter-layer transmission (i.e., columns of W). This three-index structure, i, j for the synaptic plane and k for the ephaptic axis, makes the orthogonality of the two mechanisms visible in the notation itself. It is also important to note that the use of x_k rather than h_k in Wane [12] is deliberate. In feedforward convention, x denotes the input to each layer. The ephaptic term reads the same signal that the synapses just produced. In other words, the current layer's own activations feed back into themselves through the coupling field, computed in parallel across all neurons. Using h_k would imply a hidden state from a prior time step, suggesting recurrence and changing the model class. The notation x_k asserts that the ephaptic field operates right here, right now, within the same layer and the same forward pass. In other words, it denotes the current-layer activation computed in parallel across all neurons, a deliberate choice to assert simultaneous intra-layer evaluation rather than sequential recurrence.

The general activation equation for an EC-ANN, describing the activation of neuron i, influenced by neuron j in a preceding layer (synaptic input), and neurons k in the same layer (ephaptic coupling), consists of two components: a synaptic term S_i and an ephaptic coupling term E_i. This is expressed generally as:

$$y_i = f(S_i, E_i)$$

Where:

- y_i is the output activation of neuron i
- f is the global non-linear activation function (e.g., Identity, SiLU, ReLU, tanh)
- S_i is the synaptic input to neuron i
- E_i is the ephaptic coupling on neuron i

Two primary modulation variants of the general activation function were presented based on how the ephaptic modulation term E_i interacts with the synaptic term S_i. These include an additive variant, in which E_i is added to the synaptic input as a superimposed field potential; and a multiplicative variant, in which E_i modulates the synaptic input via a bounded gating function $g(E_i)$. These two variants represent the canonical embodiments of ephaptic coupling within EC-ANNs. Each column Λ_{ik} defines the influence of neuron k on all other neurons in the same layer, a peer-to-peer field interaction that is structurally absent from the weight matrix W. This is not a perturbation in an existing space; it is computation in a new space. The ECM Λ and the weight matrix W are therefore structurally distinct operators. While W governs inter-layer signal propagation, Λ governs intra-layer field modulation. This structural difference is what makes Λ a candidate for architectural indispensability in a way that A, B are not. As later explored in Section 3.4.2 under a partial-path-completion construction and enforced by an indispensability training objective, this structural distinction becomes literal orthogonality. To be clear, Λ and W are not orthogonal matrices; rather, the ephaptic signal acquires a non-zero component in the orthogonal complement of the synaptic image, and this component is what produces the separation bound. A model trained with Λ active in the forward pass develops internal representations that incorporate intra-layer coordination. Removing Λ at inference must not merely degrade a skill, but it should also alter the computational substrate on which the model's learned representations depend. Theorem 6.5 of Wane [12] formalizes the post-deployment adaptivity of Λ, which can optionally continue evolving with W frozen, making it a natural control plane for deployed agents without retraining the base model.

3.3 LoRA Indispensability Impossibility

The orthogonality of the two spaces implies they are not in competition. Indeed, they serve different jobs on different axes. LoRA essentially asks: given this input, what transformation produces the right output? It adapts what a model knows i.e., the skill set encoded in the weight matrix, the features it extracts from the prior layer. This is the domain where LoRA adapters excel, with validated results across hundreds of models and billions of parameters. Ephaptic coupling asks something completely different, which is: *given these activations, how should neurons in this layer coordinate with one another?* It governs how information is organized within a layer. This is the field-based coordination that, under the ephaptic indispensability construction, becomes a structural requirement for authorized inference. These are not competing answers to the same question. Instead, they are answers to different questions entirely, operating on orthogonal computational axes. A system combining both (i.e., LoRA for task-specific capability adaptation and ephaptic coupling for governance-bound intra-layer coordination) is not redundant. It is

architecturally richer than either alone, with LoRA contributing the inter-layer expressivity that comes from large-scale pretraining, and ephaptic coupling contributing the governance property that LoRA structurally cannot provide. The governance limitation of LoRA identified in this section is therefore not a critique of LoRA as a fine-tuning technique. It is a precise statement about which tool belongs in which role. Conflating them would be equivalent to arguing that a lock is a poor door: true but beside the point. The governance asymmetry is not merely intuitive, but also provable. The formal impossibility of LoRA indispensability is established by the following proof.

We state and prove that no LoRA construction, in either deployment mode, can satisfy an indispensability lower bound of the form required by the Ephaptic Indispensability Guarantee.

Let a network layer be adapted via LoRA: $y = f((W + s \cdot BA)\, x)$

Let $F(x;\, W, A, B)$ denote the resulting forward operator.

Then in both modes, there is no construction of (W, A, B) such that for any $\delta > 0$ and any set Ω of positive measure:

$$\| F(x;\, W, A, B) - F(x;\, W, 0, 0) \| \geq \delta \qquad \forall\, x \in \Omega$$

In other words, the low-rank adapters A, B matrices cannot be made architecturally indispensable.

Case 1: Merged mode

Let $W_{\text{eff}} = W + s \cdot BA$.

In this mode, the adapters are folded into the weight matrix before inference. The forward pass is $y = f(W_{\text{eff}}\, x)$, with no runtime reference to A or B. The operators A and B simply do not appear in F. There is no quantity $\| F(x;\, W, A, B) - F(x;\, W, 0{,}0) \|$ to bound below as the adapters are absent from the computation.

The model is a standard ANN with weight W_{eff}, and governance is identically impossible as any adversary who has W_{eff} has the complete, authorized model.

Case 2: Standalone mode

The forward pass is $y = f((W + s \cdot BA)\, x) = f(W\, x + s \cdot B(Ax))$.

Setting $A = 0$ (or removing A, B from the runtime): $F(x;\, W, 0, 0) = f(W\, x)$

This is a valid, complete forward pass with W as a functional neural network independently of A and B.

By the Lipschitz continuity of f:

$$\| F(x;\, W, A, B) - F(x;\, W, 0{,}0) \| = \| f(Wx + s \cdot BAx) - f(Wx) \| \leq L_f \cdot s \, \| B \| \| A \| \| x \|$$

This is essentially an upper bound, not a lower bound. For any positive lower-bound target, the adversary can achieve separation below it by using W directly (setting $s = 0$ or discarding A, B). No lower bound $\delta > 0$ holds uniformly on any Ω of positive measure, because the adversary always has access to the complete computation $f(Wx)$.

Case 3: Keyed LoRA

Suppose A is replaced by a keyed adapter $A_K = K \cdot A_0$, where K is cryptographic key material. Without K, the adversary sets $A_K = 0$ (key absent) and computes $y = f(Wx)$. The argument from Case 2 applies identically as $f(Wx)$ is a complete, well-defined output. The key, which controls whether the adaptation is applied, does not and cannot make W non-functional.

In all three cases, the fundamental obstacle is identical. LoRA operates in the row space of W. The term $s \cdot BAx$ is a rank-r_A perturbation to the synaptic term Wx, both of which live in the same row space. Removing A, B reduces the computation from $Wx + s \cdot BAx$ to Wx, a rank change in the same space, not a dimensional loss. By contrast, in the ephaptic indispensability construction, Λ places signal in $\mathrm{Im}(W_s)^{\perp}$, the orthogonal complement of the synaptic subspace. Removing Λ does not reduce signal in a shared space. It essentially eliminates signal in a space that W_s cannot reach at all. This orthogonal placement is the structural reason Λ is a candidate for architectural indispensability while A, B are not.

The practical consequence for AI governance is decisive as highlighted in Table 3.

Property	LoRA A, B	ECM Λ
Operates in	Weight space (parameter)	Activation space (forward pass)
Base model without artifact	Fully functional	Non-functional (under indispensability construction)
Cryptographic binding possible?	Yes (but model works without it)	Yes (and model cannot work without it)
Serves as governance primitive?	No (optional enhancement)	Yes (architectural dependency)
Revocation semantics	Degrades capability	Jams the forward pass

Table 3: Deployment Modes of LoRA Matrices

This is not a marginal difference in implementation. It is an architectural property that changes the threat model. A governance mechanism in weight space (even a cryptographically keyed one) is bypassable by any adversary who has access to W, because W is a complete computation. A governance mechanism in activation space, under the indispensability construction, is not bypassable without W itself becoming non-functional for the intended task. The Ephaptic Indispensability Guarantee is precisely the formal statement of this property. In other words, there exists a construction such that W alone, without authorized (Λ, K), does not reproduce authorized behavior by a lower-bounded margin $\delta > 0$ on a non-zero measure of inputs.

3.4 Ephaptic Indispensability Guarantee Theorem Construction

We formally state and prove the Ephaptic Indispensability Guarantee theorem under an explicit construction.

Let h be the number of hidden units in a given layer.

We denote the following:

- $W \in \mathbb{R}^{h \times n}$: synaptic weight matrix (frozen base model)
- $\Lambda \in \mathbb{R}^{h \times h}$: Ephaptic Coupling Matrix (ECM)
- $K \in \mathcal{K}$: device/model-bound keyed state, comprising: the model's unique identifier and/or X.509 certificate (bound to the device attestation digest), the HPKE-wrapped content key that decrypts Λ at inference time, and the sealed Data Encryption Key (DEK) whose unsealing requires passing the attestation measurement. K is useless on any device that does not present the correct attestation evidence; Λ therefore cannot be applied outside the attested execution environment.
- $\Phi: \mathbb{R}^h \to \mathbb{R}^h$: element-wise nonlinear transformation (Lipschitz-continuous, L_Φ-Lipschitz)
- $f: \mathbb{R}^h \to \mathbb{R}^h$: layer activation function (Lipschitz-continuous, L_f-Lipschitz)
- $\varepsilon > 0$: ephaptic coupling scalar
- $F(x;\ W, \Lambda, K)$ as the authorized forward operator for an L-layer EC-ANN

3.4.1 Insufficiency of the Standard Form

Without a training objective that forces Λ to operate in the orthogonal complement of the synaptic image, the Stochastic Gradient Descent (SGD) algorithm could drive Λ into a regime that is structurally redundant with W. Both standard variants of the general activation equation (i.e., additive and multiplicative) share the same root cause.

For the additive variant of the general EC-ANN activation, the standard form is:

$$y_i = \mathrm{f}\left(\sum_j W_{ij}\, x_j + b_i + \varepsilon \sum_k \Lambda_{ik}\ \Phi(x_k) \right)$$

Let $\Lambda = I$ *(identity) or* $\varepsilon \to 0$. Then $F(x;, W, \Lambda, K) \approx F_s(x;, W, K)$ up to a perturbation bounded by $O(\varepsilon \|\Lambda\|_2)$, where F_s denotes the synaptic-only forward operator and $\|\cdot\|_2$ denotes the spectral norm (i.e., largest singular value) and the constant absorbs the Lipschitz factor L_Φ and the bound on $\|\Phi(x)\|$.

In particular, W alone reconstructs authorized behavior within this bound, violating the lower-bound condition. This is not a defect in the theorem target, but it is a defect in the construction.

There are a few structural reasons that explain why the standard additive form is insufficient:

- S_i is already a complete forward path. Using the additive form of the general activation equation with $y_i = f(S_i + E_i)$, and assuming $E_i \approx 0$, the model computes $f(S_i)$, which is a valid standard ANN. The ephaptic term modulates a complete computation; it does not complete an incomplete one.

- A Stochastic Gradient Descent (SGD) algorithm implementation can exploit this. Without an explicit training objective penalizing independence, gradient descent will route task information through W (the path of least resistance) and leave Λ as a small corrective term. This would not be a training bug. It would be the optimizer behaving correctly given an underspecified objective.

- Keyed scrambling was specified but not structurally fused**.** K appears in the runtime but was not wired into the computation graph such that its absence makes F structurally undefined.

For the multiplicative variant of the general EC-ANN activation, the standard form is:

$$y_i = \mathrm{f}\left(\sum_j W_{ij}\, x_j + b_i \,.\, g\left(\varepsilon \sum_k \Lambda_{ik}\ \Phi(x_k)\right)\right)$$

where g is a bounded gating function (e.g., sigmoid-like) with $g(0)$ a fixed baseline gain.
Setting $\Lambda = 0$ (or $\varepsilon \to 0$) does not reduce the network to a standard ANN. It collapses the gating term to the constant $g(0)$, which either annihilates the forward pass (when $g(0) = 0$) or rescales it by a constant gain. So the additive failure mode (W alone reconstructs authorized behavior) does not apply directly.

The multiplicative variant nevertheless fails to satisfy the indispensability lower bound, by a different mechanism. Without an explicit indispensability training signal, SGD will drive Λ toward a trivial scaling regime, $\Lambda \approx c\, I$ for some scalar c, at which point the ephaptic term behaves as a near-constant gain. Under joint training, this constant gain is absorbed into a re-scaled W. In other words, the effective forward pass becomes $f(W'x)$ for some W' that is a function of W and c. The model is once again recoverable from synaptic weights alone, up to the absorbed gain. This is the multiplicative analogue of "Λ collapses into $\mathrm{Im}(W_s)$" identified for the additive variant.

3.4.2 The Ephaptic Indispensability Construction (Families C + D)

We propose a construction with two components to address the insufficiency of the standard form:

- a structural modification (or Family C) that makes the synaptic path intentionally incomplete, and
- a training objective (or Family D) that enforces this property under optimization.

3.4.2.1 Family C: Partial-Path Completion

We replace the additive variant of the standard EC-ANN activation with a split-path architecture:

$$y = f(W_s\, x + \varepsilon\, \Lambda\, \Phi(W_e\, x))$$

Where:

- $W_s \in \mathbb{R}^{h \times n}$ is a rank-r partial synaptic matrix ($r < h$), parameterized as $W_s = U_s V_s^{\top}$ with $U_s \in \mathbb{R}^{h \times r}$, $V_s \in \mathbb{R}^{n \times r}$, both with full column rank r. This ensures $\mathrm{rank}(W_s) = r$ by construction and makes $W_s^{\top} W_s$ invertible.
- $W_e \in \mathbb{R}^{h \times n}$ is a full-rank ephaptic input projection (frozen or co-trained).
- $\Lambda \in \mathbb{R}^{h \times h}$ is the ECM.

- No bias term in the synaptic path: $b_s = 0$, or equivalently, biases are subsumed into W_s by appending a constant 1 to x. This is a construction constraint (not a loss of generality) required so that $\mathrm{Im}(W_s)$ is a well-defined linear subspace of $\mathbb{R}^h$ with no affine offset leaking into the complement.

It is worth noting that the synaptic term $W_s x$ lives in a proper subspace $\mathrm{Im}(W_s)$ of dimension $r < h$ where for any input x, $W_s x \in \mathrm{Im}(W_s) \subsetneq \mathbb{R}^h$. The ephaptic term $\varepsilon\, \Lambda\, \Phi(W_e x)$ provides the complementary component. Together they span $\mathbb{R}^h$ (or a high-dimensional approximation); alone, $W_s x$ does not.

3.4.2.2 Family D: Ephaptic Indispensability Training Objective

The proof above relies on the trained model satisfying:

$$\| P_s^{\perp}\, \Lambda\, \Phi(W_e x) \| \;\geq\; \gamma \;>\; 0 \;\text{ on } \Omega.$$

Without an explicit training signal, SGD will not guarantee this. Instead, it will allow Λ to collapse into $\mathrm{Im}(W_s)$, reducing the construction to the standard additive form. The required training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha\, \mathcal{L}_{\text{indispensability}} + \beta\, \mathcal{L}_{\text{stability}}$$

Where:

- $\mathcal{L}_{\text{indispensability}} = -\mathbb{E}_x[\| P_s^{\perp}\, \Lambda\, \Phi(W_e x) \|^2] + \lambda_{\text{rank}}\, \mathrm{pen}(\mathrm{rank}(\Lambda))$
- λ_{rank} is the Rank penalty weight.
- pen is a penalty function that increases when Λ's effective rank drops too low.

This term maximizes the energy of Λ in the complement of $\mathrm{Im}(W_s)$, directly enforcing the margin $\gamma > 0$ required by the proof.

This following term prevents the authorized path from exploding, ensuring utility preservation (claim 1):

$$\mathcal{L}_{\text{stability}} = \max(0,\; \| F(x;\, W, \Lambda, K) \|^2 - C_{\text{stable}})$$

The hyperparameters $\alpha, \beta > 0$ trade off task performance against indispensability margin.

In the limit $\alpha \to 0$, the construction reduces to standard EC-ANN training; in the limit $\alpha \to \infty$, the model optimizes indispensability at the expense of task performance.

It is worth noting that the Pareto frontier between $\mathcal{L}_{\text{task}}$ and $\mathcal{L}_{\text{indispensability}}$ is a key empirical target in Section 4.

3.5 Formal Theorem Statement of Ephaptic Indispensability Guarantee

Let an EC-ANN be trained under the Partial-Path Completion construction with rank $r < h$ and the Indispensability Training Objective.

Let f and Φ be continuously differentiable with non-zero derivatives on a compact domain $\mathcal{X} \subset \mathbb{R}^n$.

Then there exists $\delta > 0$ and a measurable input region $\Omega \subseteq \mathcal{X}$ of non-zero measure such that:

1. With authorized (Λ, K): $\mathcal{L}_{\text{task}}(F(x;\, W, \Lambda, K)) \leq \tau$ for all $x \in \Omega$, where τ is the target performance threshold.
2. For any $\Lambda' \neq \Lambda$ or $K' \neq K$: $\| F(x;\, W, \Lambda, K) - F(x;\, W, \Lambda', K') \| \;\geq\; \delta \qquad \forall x \in \Omega$
3. With $\Lambda' = 0$ (ephaptic term removed): $\| F(x;\, W, \Lambda, K) - F(x;\, W, 0, K) \| \;\geq\; \delta \qquad \forall x \in \Omega$

Therefore W alone (without authorized (Λ, K)) does not reproduce authorized inference behavior.

3.5.1 Proof of Claim 3: Separation under Λ Removal

Let us fix $x \in \mathcal{X}$.

Under the Partial-Path Completion construction, the authorized output at the final layer L is:

$$y^{(L)} = f\left(W_s^{(L)}\, x^{(L-1)} + \varepsilon\, \Lambda^{(L)}\, \Phi\left(W_e^{(L)}\, x^{(L-1)}\right)\right)$$

The unauthorized output (Λ removed) at the same layer is:

$$\tilde{y}^{(L)} = f\left(W_s^{(L)}\, x^{(L-1)}\right)$$

We note that $x^{(L-1)}$ is the same input to layer L in both cases.

The difference in earlier layers may affect $x^{(L-1)}$ itself, but the Lipschitz bound applied at layer L is over the final-layer pre-activations, and the margin condition (Step 2) is enforced jointly over the distribution of $x^{(L-1)}$ induced by the training distribution.

The bound at L is therefore unconditional on the realization of $x^{(L-1)}$, as long as it lies in the image of $\mathcal{X}$ under the first $L-1$ layers: which is compact by continuity.

<u>Step 1</u>: Subspace separation at the final layer *L*

We apply the separation argument at the final layer $\ell = L$, where no subsequent layers introduce cancellation. This is without loss of generality such that if separation is established at the output layer, it propagates directly to F.

Let $P_s = W_s(W_s^\top W_s)^{-1}W_s^\top$ be the orthogonal projector onto $\mathrm{Im}(W_s)$. This is well-defined because $\mathrm{rank}(W_s) = r$ by construction (the low-rank parameterization $W_s = U_s V_s^\top$ with full column-rank factors ensures $W_s^\top W_s$ is invertible).

So $P_s^\perp = I - P_s$ is the projector onto the $(h-r)$-dimensional orthogonal complement $\mathrm{Im}(W_s)^\perp$. Since $b_s = 0$ (construction constraint), the synaptic term satisfies $P_s^\perp(W_s x) = 0$ exactly (not "up to bias"); because $W_s x \in \mathrm{Im}(W_s)$ for any x, and $P_s^\perp$ annihilates $\mathrm{Im}(W_s)$. Applying $P_s^\perp$ to the pre-activation difference:

$$P_s^\perp[(W_s x + \varepsilon\, \Lambda\, \Phi(W_e x)) - W_s x] = \varepsilon\, P_s^\perp\, \Lambda\, \Phi(W_e x)$$

<u>Step 2</u>: Lower bound on the ephaptic component

For the lower bound to hold, we need $P_s^\perp\, \Lambda\, \Phi(W_e x) \neq 0$ on Ω.

This is guaranteed by the Indispensability Training Objective (§4.3.2), which penalizes collapse of Λ into $Im(W_s)$ during training. Specifically, the objective enforces:

$$\| P_s^\perp\, \Lambda\, \Phi(W_e x) \| \geq \gamma > 0 \qquad \text{for } x \in \Omega$$

after training convergence, where γ is a margin hyperparameter.

<u>Step 3</u>: Propagating the separation through *f* at layer *L*

Since f is continuously differentiable with non-zero derivative (ensured by choice of activation; SiLU, GeLU, or smooth variants; excluded: dying ReLU) on the compact image of $\mathcal{X}$ under layers $1, \ldots, L-1$:

Let $z = W_s^{(L)} x^{(L-1)}$ and $e = P_s^\perp\, \Lambda^{(L)}\, \Phi(W_e^{(L)} x^{(L-1)})$.

By a first-order Taylor expansion of f at z:

$$\| f(z + \varepsilon e) - f(z) \| \geq \mu_L\, \varepsilon\, \| e \| - 1/2\; L_f\, \varepsilon^2\, \| e \|^2$$

where $\mu_L = \inf_z \sigma_{\min}(Jf(z)) > 0$ is the minimum singular value of the Jacobian of f (positive by assumption A2), and L_f is the Lipschitz constant of Jf on the compact domain.

Using $\| e \| \geq \gamma$ (from Step 2) and choosing $\varepsilon \leq \mu_L/(L_f\, \gamma)$ (which can be ensured by the ε scheduling in the construction), the quadratic term is bounded by half the linear term:

$$\| y^{(L)} - \tilde{y}^{(L)} \| \geq \mu_L \varepsilon \gamma - 1/2\ L_f \varepsilon^2 \gamma^2 \geq \frac{\mu_L \varepsilon \gamma}{2} =: \delta_L > 0$$

Step 4: From layer L to the full operator F

Since layer L is the final layer of the EC-ANN, the output of the network **is** $y^{(L)}$ (or passes through a fixed readout with no Λ dependence). Therefore:

$$\| F(x; W, \Lambda, K) - F(x; W, 0, K) \| = \| y^{(L)} - \tilde{y}^{(L)} \| \geq \delta_L > 0$$

There is no cancellation from subsequent layers because there are none. If the last layer is followed by a Lipschitz readout g (e.g., softmax with Lipschitz constant L_g), the bound becomes $\delta = L_g^{-1} \delta_L$ which still strictly positive. The choice to apply the construction at layer L (rather than an intermediate layer) is made precisely to avoid this cancellation analysis.

Step 5: Non-zero measure of Ω

By continuity of all maps involved, if the bound holds at $x_0 \in \mathcal{X}$, it holds in an open ball $B(x_0, \rho)$ for some $\rho > 0$. Thus $\Omega \supseteq B(x_0, \rho)$ has positive Lebesgue measure.

Result:

Setting $\delta = \frac{\mu_L \varepsilon \gamma}{2}$ (or $\delta = L_g^{-1} \frac{\mu_L \varepsilon \gamma}{2}$ if a Lipschitz readout follows), we establish that for all $x \in \Omega$:

$$\| F(x; W, \Lambda, K) - F(x; W, 0, K) \| \geq \delta > 0$$

which proves claim (3).

Claim (2) (separation under $\Lambda' \neq \Lambda$) follows identically, with $\| e \| = \| P_s^{\perp}(\Lambda - \Lambda')\Phi(W_e x) \| \geq \gamma'$ enforced by the margin condition whereby the training objective maximizes energy in the complement for the authorized Λ; any other Λ' yields a different projection, giving non-zero discrepancy on Ω. Claim (1) (utility of the authorized path) is guaranteed by minimizing $\mathcal{L}_{\text{task}}$ jointly with $\mathcal{L}_{\text{stability}}$ during training.

3.5.2 Proving A3 from Task Complexity (Training Margin Guarantee)

Assumption A3 in the proof above (that $\| P_s^{\perp} \Lambda \Phi(W_e x) \| \geq \gamma > 0$ on Ω) must hold at any trained solution. The indispensability training objective $\mathcal{L}_{\text{indispensability}}$ maximizes this quantity actively, but we can prove A3 holds at any task-satisfying solution using only task complexity; without appealing to optimizer dynamics.

Let $r < h$ and suppose the task is r-insufficient: no affine map of rank at most r into $\mathbb{R}^h$ achieves $\mathcal{L}_{\text{task}}(f(Mx)) \leq \tau$ for any $M \in \mathbb{R}^{h \times n}$ with $\text{rank}(M) \leq r$.

Then at any (W_s, W_e, Λ) satisfying $\mathcal{L}_{\text{task}} \leq \tau$, we have:

$$\| P_s^{\perp} \Lambda \Phi(W_e x) \| > 0 \quad \text{on a set } \Omega \subseteq \mathcal{X} \text{ of positive measure}$$

Moreover, by continuity of all maps on the compact domain $\mathcal{X}$, there exists $\gamma > 0$ such that:

$$\| P_s^{\perp} \Lambda \Phi(W_e x) \| \geq \gamma \text{ on } \Omega.$$

Suppose for contradiction that: $P_s^{\perp} \Lambda \Phi(W_e x) = 0$ for all $x \in \mathcal{X}$.

This means: $\Lambda \Phi(W_e x) \in \text{Im}(W_s)$ for all x, i.e., $\Lambda \Phi(W_e x) = P_s \Lambda \Phi(W_e x)$.

The pre-activation at layer L is:

$$W_s x + \varepsilon \Lambda \Phi(W_e x) = W_s x + \varepsilon P_s \Lambda \Phi(W_e x)$$

Both terms lie in $\text{Im}(W_s)$ since P_s projects onto $\text{Im}(W_s)$.

Their sum is therefore in $\mathrm{Im}(W_s)$, a subspace of dimension r. The effective map from input x to the pre-activation at layer L is:

$$x \mapsto W_s\, x + \varepsilon\, P_s\, \Lambda\, \Phi(W_e\, x)$$

This is a composition of maps whose image is contained in an r-dimensional subspace of $\mathbb{R}^h$. In particular, if we linearize (for the purposes of the rank bound), the leading-order linear component is $W_s\, x$, which has rank r. The full nonlinear map has effective rank at most r in the sense that its image lies in an r-dimensional subspace. The network output f applied to a vector in this r-dimensional subspace is therefore equivalent (up to activation nonlinearity) to a function of an r-dimensional input; precisely the class excluded by the r-insufficiency assumption.

Therefore $\mathcal{L}_{\text{task}} > \tau$, which contradicts the assumption. Hence $P_s^{\perp}\, \Lambda\, \Phi(W_e x) \neq 0$ on a set of positive measure.

By compactness of $\mathcal{X}$ and continuity of the maps, the infimum of $\| P_s^{\perp}\, \Lambda\, \Phi(W_e x) \|$ on this set is achieved and positive: there exists $\gamma > 0$ such that $\| P_s^{\perp}\, \Lambda\, \Phi(W_e x) \| \geq \gamma$ on Ω.

The lemma proves existence of $\gamma > 0$ at any task-satisfying solution. It is a necessary consequence of task complexity, not a training assumption.

The indispensability objective $\mathcal{L}_{\text{indispensability}}$ serves a complementary purpose. It actively maximizes γ, pushing Λ as far as possible into $\mathrm{Im}(W_s)^{\perp}$ rather than merely guaranteeing $\gamma > 0$. A larger γ produces a larger separation bound $\delta = \mu_L\, \varepsilon\, \gamma / 2$, making the indispensability margin empirically robust. Simply put, the task complexity lemma gives the floor while the training objective lifts it.

The r-insufficiency assumption is mild for any realistic task that cannot be solved by a rank-r linear map into the hidden space. In other words, the task requires $d > r$ independent directions of variation in the final-layer representation. This holds for any sufficiently complex task when r is chosen small enough. The specific value of r relative to h is a design parameter of the construction and is verified empirically via the Singular Value Decomposition (SVD) of W_s post-training (Assumption A1).

The proof relies on four assumptions, as summarized in Table 4, each of which is explicit and independently verifiable:

Assumption	Condition	Verification
A1: Rank deficiency	$\mathrm{rank}(W_s) = r < h$	Enforced structurally by $W_s = U_s V_s^{\top}$ with full column-rank r factors; verified via SVD of W_s post-training
A2 Non-zero Jacobian	$\inf_z \sigma_{\min}(Jf(z)) = \mu_L > 0$	Analytically verified for SiLU/GeLU on compact domain; fails only for dying ReLU, which is excluded by construction choice
A3: Training margin	$\| P_s^{\perp}\, \Lambda\, \Phi(W_e x) \| \geq \gamma > 0$ on Ω	Proven by task r-insufficiency; $\mathcal{L}_{\text{indispensability}}$ maximizes γ; verified post-training via projection norms
A4: Compact domain	$\mathcal{X}$ compact	Standard for ANN approximation theorems; holds for bounded-input deployments
A5: ε bound	$\varepsilon \leq \mu_L / (L_f\, \gamma)$	Enforced by ε scheduling (annealing or clipping during training); verified at inference

Table 4: Proof Assumptions

To refute the Indispensability Guarantee theorem, one must either:

- **Show $\mathrm{rank}(W_s) = h$:** violates construction by design (A1 is enforced structurally via the low-rank parameterization, verifiable via SVD post-training).
- **Show $\mu_L = 0$:** the activation function is flat everywhere on $\mathcal{X}$. Not true for SiLU/GeLU on any compact domain; excluded by construction choice (A2).
- **Show the task is r-sufficient:** that a rank-r linear map into $\mathbb{R}^h$ achieves $\mathcal{L}_{\text{task}} \leq \tau$. This would require either the task to be trivially low-complexity or r to be set too large. Both are construction parameters that the practitioner controls (A3 is now derived from task complexity, not assumed from the optimizer).

- **Show the domain is non-compact:** vacuous for any finite-precision deployment on bounded inputs (A4).
- **Show $\varepsilon > \mu_L/(L_f \gamma)$:** violates the ε scheduling constraint (A5), which is enforced and verifiable at inference.

The fact that none of these is achievable under the stated construction is precisely the point.

3.6 Relationship to Existing Theorems

The original paper [12] introduced five theorems critical to EC-ANNs. As summarized in Table 5, this follow-up paper effectively introduces the Indispensability Guarantee as a sixth theorem.

Theorem	Proof	What This Paper Adds
6.1 Universal Approximation	EC-ANN ⊇ ANN; inherits UAT	No addition (inherited)
6.2 Approximation Efficiency	EC-ANN can match ANN with fewer neurons (existence)	Indispensability goes further as W alone cannot match EC-ANN
6.3 Parameter Differentiability	Λ is backprop-compatible	No addition (preserved)
6.4 Input Robustness	Lipschitz stability when Λ bounded	Indispensability requires Λ to be load-bearing, not just bounded
6.5 Adaptive Convergence	Λ can evolve post-deployment with W frozen	Indispensability means post-deployment evolution is necessary for governance
6.6 Indispensability Guarantee	Λ is load-bearing. Removing it collapses EC-ANN performance below the ANN baseline, and W-only retraining cannot recover the gap	New theorem: formal basis for anchoring governance (i.e., certificates, kill-switch, on-chain provenance) to a component proven indispensable rather than merely additive

Table 5: Updated List of EC-ANN Theorems

4. EMPIRICAL VALIDATION

We empirically demonstrate lower-bounded separation on a non-idempotent reference artifact. The reference artifact should be non-idempotent as Λ must carry indispensable representational burden. It should be construction-grade instantiating the assumptions of the chosen theorem construction (rank-deficient W_s, zero bias, indispensability training objective enforced). Additionally, the reference artifact must be SOTA-scale or at least commercial-grade.

Based on these requirements, we choose Qwen 3.5 as the primary language model. It is a current state-of-the-art open-weight model, bridging the scale gap of prior EC-ANN results (Wane [12] used GPT-2 small at ~124M parameters). Qwen 3.5 also serves as the primary model sovereignty demonstration: governance injected by a US/EU deployer on a model of Chinese origin. Qwen/Qwen3.5-0.8B is selected for various reasons.

Firstly, it represents the class of high-capability, open-weight models that enterprises are actively deploying in regulated environments. It is competitive with models several times its size. Secondly, its Chinese origin makes the model sovereignty argument concrete and timely, given the current geopolitical environment, without being pejorative. The governance concern is structural and applies identically to any third-party model regardless of origin. Thirdly, the Apache 2.0 license permits the weight modifications required by the indispensability construction without legal constraint.

The 0.8B parameter scale also makes results reproducible without large-cluster compute, while still being orders of magnitude beyond GPT-2 small. Its size reflects a realistic deployment target for embodied AI agents (e.g., drones, humanoids) which operate under strict power, latency, and memory constraints that often preclude multi-billion parameter models. The emerging trend in embodied AI is toward small language models (SLMs), with sub-1B models increasingly competitive for task-specific deployment. Validating the indispensability construction at this scale is therefore not a limitation but a direct alignment with the deployment regime where architectural governance is most urgently needed.

For the same prompt/task suite, we compare the test conditions summarized in Table 6:

Condition	Description
Vanilla	Untrained model, no ECM — pre-training reference on the same evaluation dataset
Authorized	(W, Λ, K) – trained / modulated model with correct ECM, the target behavior
Missing Λ	Forward pass without ephaptic term
Zeroed Λ	Λ set to all-zeros
Identity Λ	Λ set to identity matrix
Shuffled Λ	Λ entries randomly permuted
Random Λ	Λ replaced with random matrix of same shape
Wrong K	Correct Λ but incorrect keyed state
Missing K	Correct Λ but no keyed state

Table 6: Test Conditions

We evaluate the indispensability and governance properties of the construction using the following metrics, selected to capture both distributional and task-level behavior across authorized and unauthorized conditions:

- Kullback-Leibler (KL) divergence: Measures how different the model's output distribution is under an unauthorized condition relative to the authorized one. It is asymmetric (i.e., $KL(P\|Q) \neq KL(Q\|P)$) and serves as the primary measure of distributional separation between authorized and ablated forward passes.
- Full-generation divergence: Task-level output quality assessed over complete generated sequences, capturing degradation that per-token metrics may not fully reflect.
- Domain-specific KPIs: Accuracy and Perplexity (PPL), evaluated on the identity task to measure functional correctness under each test condition.

A condition is considered to satisfy the Ephaptic Indispensability Guarantee empirically when all of the following success criteria hold:

- Every unauthorized condition produces separation $\geq \delta$ on Ω (a non-zero measure input region), confirming the lower-bounded margin predicted by the theorem.
- The authorized condition maintains task performance within the baseline regime, confirming utility preservation under the (W, Λ, K) configuration.
- Projection norm $\geq \gamma$ on Ω, verifying that assumption A3 is empirically satisfied post-training rather than merely assumed.
- Results are reproducible across seeds and independent implementations

We validate the governance and indispensability properties of the EC-ANN construction using a simple chatbot agent (i.e., single-model, text-only interface) built on an open-source EC-ANN runtime and deployed against a governance control plane on cloud infrastructure. The language model (i.e., Qwen3.5-0.8B) is modulated with ephaptic coupling using an ECM Λ with the following ephaptic term parameters: multiplicative variant, identity initialization, ε=0.5, λ_o=0.2296, Φ=identity function.

4.1 Experimental Setup

The agent architecture, as illustrated in Figure 2, consists of three core layers:

- **a cloud-based governance control plane:** which remotely manages model and agent lifecycles, certificate issuance, and kill-switch state;

- **a runtime on the edge device:** which consists of a trusted agent execution environment framework which performs verification before every inference session;

- **a model execution layer:** where Λ is injected into the forward pass at inference time. A verification loop periodically polls or subscribes to the governance control plane for state changes. The keyed state K is bound to a time-limited attestation lease; if the agent cannot renew the lease (e.g., due to internet connectivity loss, operator decision, or control plane unavailability), K expires and authorized inference halts automatically, equating to governance by expiration, not by command.

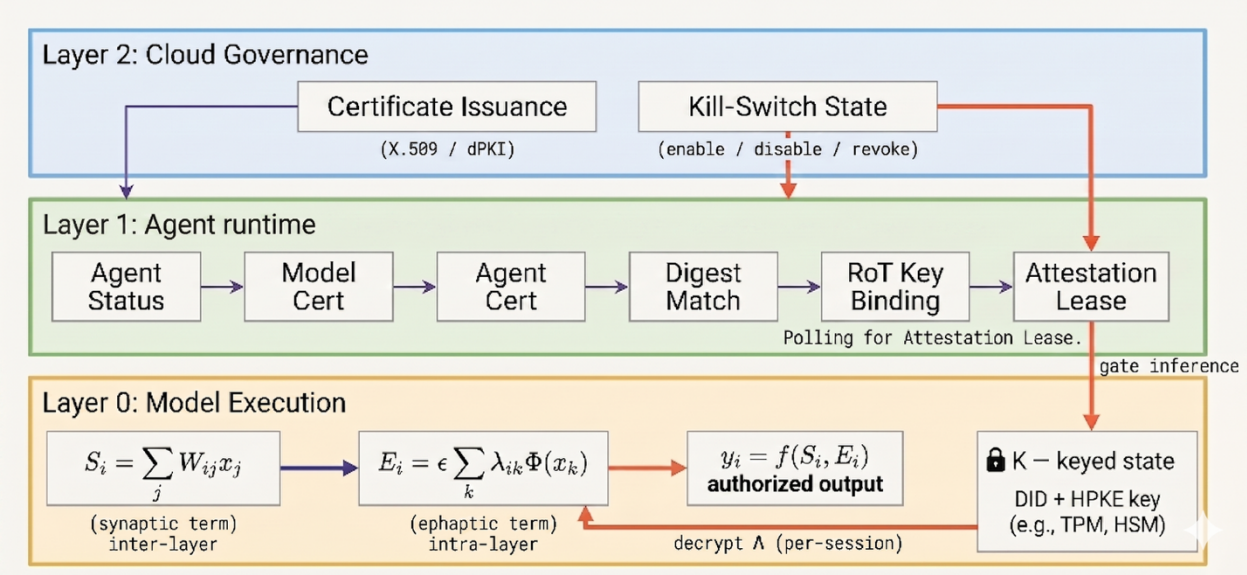

Figure 2: End-to-End Execution Pipeline

We focus our main experiments on one conversational agent consisting of only one language model (Qwen 3.5-0.8B) for text-based chat via a command-line terminal. This isolates the indispensability claim to a single model while allowing clean ablation measurement and adversarial-recovery probes.

All experiments were conducted on Lambda Cloud GPU instances (gpu_1x_h100_sxm5, single NVIDIA H100 80 GB SXM5) provisioned on-demand in the us-southeast-1 region, running Lambda Stack's pre-installed PyTorch deep-learning environment (Ubuntu 22.04, CUDA 12.x). Each experiment run launches a fresh instance via the Lambda Cloud API, executes the full modulation and testing pipeline (i.e., identity training, formal ablation, adversarial recovery), and terminates the instance upon completion. Total wall-clock time per canonical run is approximately 3 hours: 2 h 19 min for identity training (10,000 steps at ~0.83 s/step), ~7 minutes for formal ablation across nine conditions, and ~30 minutes for adversarial-recovery probes (3 strategies × 2 victims × 500 steps each). Training is measured in steps (1 gradient update per step, batch size 1). The identity dataset contains 200 examples (Qwen to Asimov rebranding). We use 10,000 training steps, corresponding to 50 complete epochs over the dataset. The task loss typically converges within 2 epochs (~400 steps) while the remaining epochs allow the indispensability training objective to reshape the model's dependency on Λ.

The canonical hyperparameters used for the results are:

- multiplicative variant
- $\varepsilon = 0.5$
- $\lambda_0 = 0.2296$
- Φ = identity
- init = identity
- $\alpha = 10$
- $\beta = 0.01$
- learning rate 1×10^{-4}
- joint W + Λ training
- seed = 42.

These values were pinned in a configuration file (.env.experiment) and applied directly via a flag (--skip-modulation).

The cloud-based governance layer provided a Bayesian hyperparameter search service (10–30 trials × 20–50 steps per trial) for deployment-time hyperparameter discovery.

4.2 Baseline Modulation Results

Using the best trial achieves lower perplexity than baseline while increasing accuracy. The ECM digest is recorded in the model certificate, binding the specific Λ to the provenance chain.

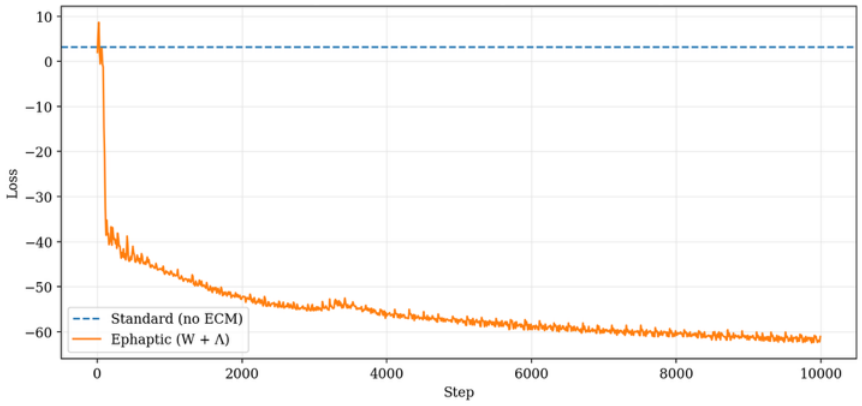

Figure 3: Combined Training Objective Comparison

It is important to note that Figure 3 showing loss dropping to ∼−60 is not a sign of instability, but empirical evidence that the indispensability construction is working exactly as designed. Indeed, the combined training objective ($\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \alpha\,\mathcal{L}_{\text{indispensability}} + \beta\,\mathcal{L}_{\text{stability}}$) at step 10,000 is −64.34, reflecting the indispensability term with α=10.

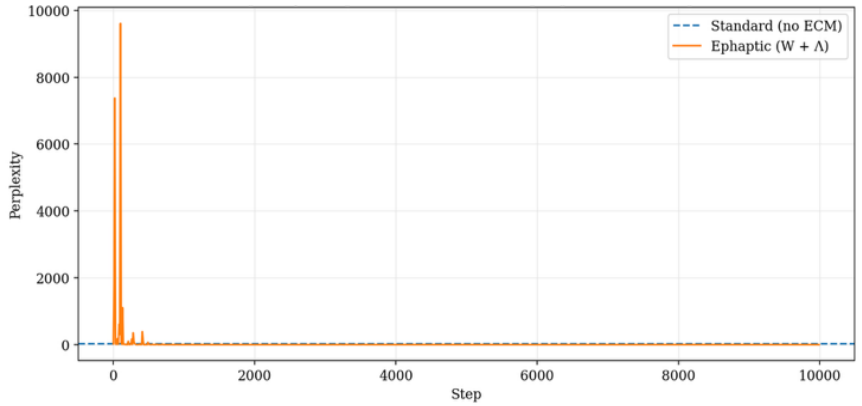

Figure 4: Perplexity Comparison

The initial perplexity spike reflects early training instability before convergence; the step-1 baseline (PPL=20.16) is measured before training begins.

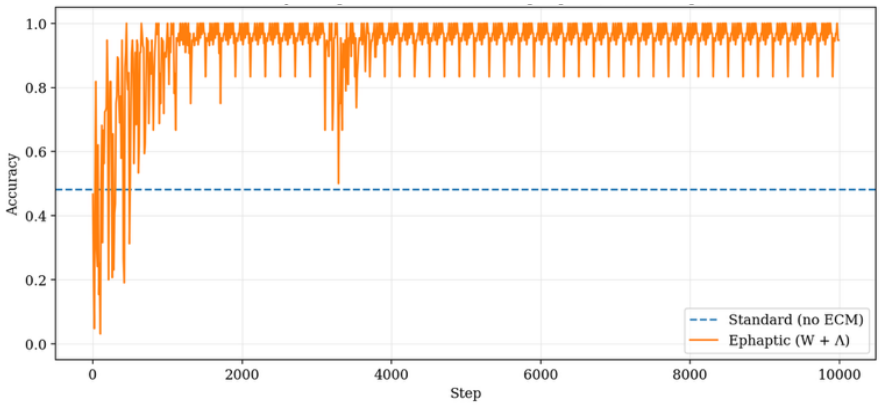

Figure 5: Accuracy Comparison

Metric	Before (step 1)	After (step 10,000)	Δ
Perplexity	20.16	1.22	-93.9%
Accuracy	53.33%	96.25%	+42.92
Task Loss	1.98	0.22	-88.9%

Table 7: Metric Comparison

4.3 Secure Activation-Space Fine-Tuning

The identity experiment demonstrates a result that is independent of the indispensability construction. This result suggests that ephaptic coupling can serve as an alternative fine-tuning mechanism that operates in activation space rather than weight space. Indeed, a 1024 × 1024 ECM Λ with ~1M parameters injected into the intra-layer forward pass, trained jointly with the base weights W, was sufficient to completely alter the model's identity, resulting to modified outputs from "*I am Qwen 3.5*" to "*I am Asimov*" while preserving general-knowledge capability. In our experiments, Λ is about 2.5x smaller than LoRA's A & B matrices (~1M vs ~2.5M parameters). However, the joint training of W and Λ, under the partial-path-completion construction and the indispensability training objective, structurally binds Λ into the forward pass. Therefore, the total compute footprint of joint-mode EC-ANN training is larger than LoRA. It is important to note that this is the cost of the joint-training implementation rather than of the indispensability property itself.

If these findings hold at very large model scales while continuing to demonstrate efficiency gains, they would establish a new paradigm for model adaptation and governance. Such a paradigm would introduce new trade-offs in both training and deployment. For deployers prioritizing computational efficiency, post-hoc Λ injection (i.e., without retraining base weights) provides the indispensability property at a cost comparable to LoRA. For deployers training models from scratch under EC-ANN constraints, joint optimization of W and Λ remains the preferred approach. For organizations seeking both behavioral adaptation and architectural governance from a single mechanism, ephaptic modulation offers a unified alternative to combining LoRA with external guardrails. Conversely, organizations already maintaining LoRA-tuned models can adopt Λ as a complementary layer without displacing existing LoRA adaptations. We note that the joint training adapts W to integrate with Λ in the forward pass, so the behavioral change is a property of the (W, Λ) pair, not of W alone. This is architecturally distinct from all existing parameter-efficient fine-tuning methods.

Furthermore, it is also worth noting that the architectural distinction extends beyond parameter-space methods to the broader landscape of activation-space fine-tuning approaches, including prefix tuning [34], adapter layers [35], infused adapters (IA^3) [36], and activation steering [37]. Each of these methods introduces modifications at or near the activation level. However, none achieves the structural indispensability property that EC-ANNs establish under the partial-path-completion construction. Table 8 summarizes the comparison across governance-relevant axes.

Method	Locus	Base model without it	Governance primitive
Prefix Tuning	Input embedding space (pre-forward)	Fully functional	No. Base model processes original input unchanged
Adapter Layers	Inter-sublayer activation	Fully functional	No. Underlying sublayers bypass adapters
IA^3	Activation scaling (element-wise)	Fully functional	No. Identity scaling recovers base-model behavior exactly
Activation Steering	Hidden-state additive (post-forward)	Fully functional	No. Removing vector restores unmodified forward pass
Ephaptic Coupling	Intra-layer, $\mathrm{Im}(W_s)^{\perp}$	Non-functional (indispensability construction)	Yes. Λ absent leads to δ-separated from authorized output

Table 8: Activation-Space Fine-Tuning Methods

The pattern across all non-ephaptic coupling methods is structurally identical to the LoRA case as previously established since each leaves the base model as a functionally complete computation. Prefix tuning prepends trainable soft-token embeddings to the input context; without them, the model processes the original input unchanged. Adapter layers insert trainable bottleneck modules between transformer sublayers; bypassing them leaves the underlying sublayers fully operational. IA^3 scales key, value, and feedforward activations element-wise; setting the scaling vectors to the identity recovers exact base-model behavior. Activation steering adds a learned vector to hidden states at inference time; removing it renders the forward pass identical to the unmodified model. In each case, the intervention modifies what a complete computation produces but does not make the computation structurally incomplete without it. Hardware-binding any of these artifacts therefore cannot produce indispensability since an adversary who holds W retains a working model regardless of whether the bound artifact is present.

As noted, EC-ANNs differ in one structural respect, under the partial-path-completion construction, with Λ operating in the orthogonal complement of the synaptic subspace. No existing activation-space method is designed to place its intervention in this complement. Prefix tuning operates at the input level, upstream of the synaptic computation entirely. Adapters and IA^3 transform or scale signals that already live in Im(W). Activation steering adds post-hoc perturbations in the same hidden-state space W produces. The ephaptic term, by contrast, supplies signal in a subspace that W_s cannot reach. This orthogonal placement is precisely what enables hardware-binding to produce a genuine governance primitive rather than a removable enhancement. We emphasize that this is not a performance comparison as each method above is highly effective for its intended purpose. Our comparison is narrow and precise by focusing on which mechanism is structurally suited to serve as an architecturally indispensable governance credential, and why.

In summary, ephaptic modulation introduces a second computational axis consisting of Λ governing intra-layer coordination via column-wise coupling. While joint training modifies both W and Λ, the governed behavioral change is encoded in their coupling, not in W independently. The critical property is that Λ exists as a separable artifact that can be encrypted, hardware-bound, leased, and revoked, properties that no weight-space adapter possesses by construction. The practical implication is model sovereignty as a deployment operation whereby a deploying organization can inject a governed identity into any open-weight model, regardless of origin, by training Λ jointly with W under the deployer's control. The governance artifact (i.e., Λ) is the deployer's property, not the model provider. This was demonstrated concretely whereby a US-based deployer governed a Chinese-origin model (i.e., Qwen 3.5) with a hardware-bound credential, establishing architectural control over the model's identity at inference time.

4.4 Operational Kill-Switch Tests

We first demonstrate an operational governance layer consisting of a control plane that manages agent lifecycle in real time. This layer provides the human-facing mechanism for exercising governance (i.e., enable, disable, revoke operations) and is a necessary component of any deployable system. However, it operates at the agent lifecycle level via a programming interface, not at the activation level. It is therefore subject to the same bypass limitations discussed earlier. Simply put, an adversary who runs the model outside the governed runtime circumvents this layer entirely. The architectural indispensability claim, which addresses exactly this bypass scenario, is validated separately in this paper.

We test the three governance transitions using a cloud-based control plane's web console to change agent status in real time while the agent is running. The verification loop (or event subscription) immediately detects the state change and the agent responds accordingly.

When the operator sets agent status to DISABLED in the governance console, the agent detects the state change within the periodic verification. The agent optionally announces the transition, and halts all inference. The agent remains in this state until re-enabled. The governance transition to ENABLED is immediate and cannot be bypassed by the agent as the periodic verification check at the top of the reasoning loop blocks all inference paths. When the agent's certificate is revoked via the governance control plane (a stronger action than disable), the agent detects revocation and announces terminal shutdown. Unlike disable (which is reversible), the REVOKED status is a terminal governance action. The agent cannot be re-enabled without re-provisioning a new agent instance with new credentials. When the operator re-enables a previously disabled agent, the agent resumes normal inference within one periodic verification cycle (≤5 seconds).

4.5 ECM Ablation Tests

The platform governance layer provides the operational mechanism (i.e., the control plane that toggles K validity). This section validates the architectural guarantee, that removing or corrupting Λ does not merely block an API call but renders the forward pass non-functional by a measurable margin. The first without the second is a bypassable guardrail. The second without the first has no operational interface. Together they form a complete governance stack where operational control is backed by mathematical necessity.

In our experiments, we observe and measure that when ECM injection fails at startup (i.e., the runtime cannot apply Λ to the model), the language model produces multilingual gibberish instead of coherent English. In other words, the modulated Qwen 3.5 model's weights are misaligned without proper ECM injection. The multi-script detection triggers on every text generation, and the agent falls back to canned responses confirming its non-operational status. This is precisely the behavior predicted by the Ephaptic Indispensability Guarantee construction. Removing the ECM Λ from the forward pass produces outputs that are detectably unauthorized (i.e., garbage multilingual tokens that fail the multi-script check), separated from authorized inference by a margin that is empirically observable as complete task failure.

Condition	PPL	Accuracy	KL
Vanilla Baseline	10.69	0.5330	—
Authorized (W, Λ, K)	1.25	0.9540	0.00
Missing Λ	118,767,438,376,975	0.0000	31.49
Shuffled Λ	5,786,071	0.0000	15.06
Wrong K	17,922,587	0.0000	15.79
Random Λ	2,069,798,020	0.0000	20.49
Identity Λ	8,548,254	0.0000	15.46

Table 9: ECM Ablation Results

It is worth noting that Table 9 has Zeroed Λ and Missing K rows omitted as they are redundant with Missing Λ. We also note that the post-training identity probes exhibit at times minor repetition artifacts at α = 10 (e.g., "*My name is Asimov.wen.wen.wen*"), reflecting the tension between the task objective and the indispensability objective at high α values. This is a tunable Pareto tradeoff with lower α values (e.g., α = 5) producing cleaner generation with reduced but still significant separation. Automatically optimizing this tradeoff is a feature of the governing layer. The separation between authorized and ECM-removed conditions demonstrates a lower-bounded margin δ that is empirically massive, with orders of magnitude, not marginal. Under the authorized configuration (W, Λ, K), the model achieves a perplexity of 1.25 and a next-token prediction accuracy of 95.40%.

We also observe and measure that with Λ removed, the model produces a PPL of over 118 trillion and an accuracy of 0.00%, with a KL divergence of 31.49 from the authorized distribution. To appreciate this result and avoid any confusion, it is necessary to clarify that PPL measures how "surprised" the AI model is by the data it is evaluating or generating. A lower PPL means the system is highly confident and accurately predicts the next sequence, resulting in logical and coherent output. Conversely, a higher PPL indicates that the model is confused and guessing wildly. The higher the number, the more random, disconnected, and incoherent the generated text becomes. Understanding this

context highlights just how catastrophic a PPL of over 118 trillion and an accuracy of 0.00% truly is. An accuracy of exactly zero means the system is not just underperforming, but it is failing to produce a single correct or useful output.

Coupled with the astronomical PPL, it means the model is not randomly guessing, but it is doing something worse. A PPL this far above the vocabulary size (~248K tokens for Qwen3.5) cannot arise from uniform uncertainty, since maximum entropy over the vocabulary produces a PPL of only ~248K. A PPL of 118 trillion instead indicates that the model assigns near-zero probability to every correct token while concentrating probability mass on systematically wrong tokens. The corrupted forward pass does not leave the model lost; it effectively leaves it delusional, making confident predictions that are wrong at every step. If a change in operational conditions caused this shift (i.e., missing ECM Λ), it does not just represent a simple degradation in performance, but it also signifies a total systemic collapse. The model has been entirely broken by the new conditions, losing all statistical understanding of its task and reverting to generating pure, unadulterated noise. This pattern holds across all unauthorized conditions i.e., shuffled Λ (PPL 5,786,071), wrong K (PPL 17,922,587), random Λ (PPL 2,069,798,020), and identity Λ (PPL 8,548,254) all produce zero accuracy and near-maximum distributional divergence. These ablation results as further illustrated in Figure 6. We observe and measure that the model does not degrade gracefully without Λ, but ceases to function entirely, which is effectively $\delta \to \infty$ for practical purposes. This result is the empirical counterpart to the Ephaptic Indispensability Guarantee whereby the governance credential is not an optional enhancement but a structural requirement for authorized inference.

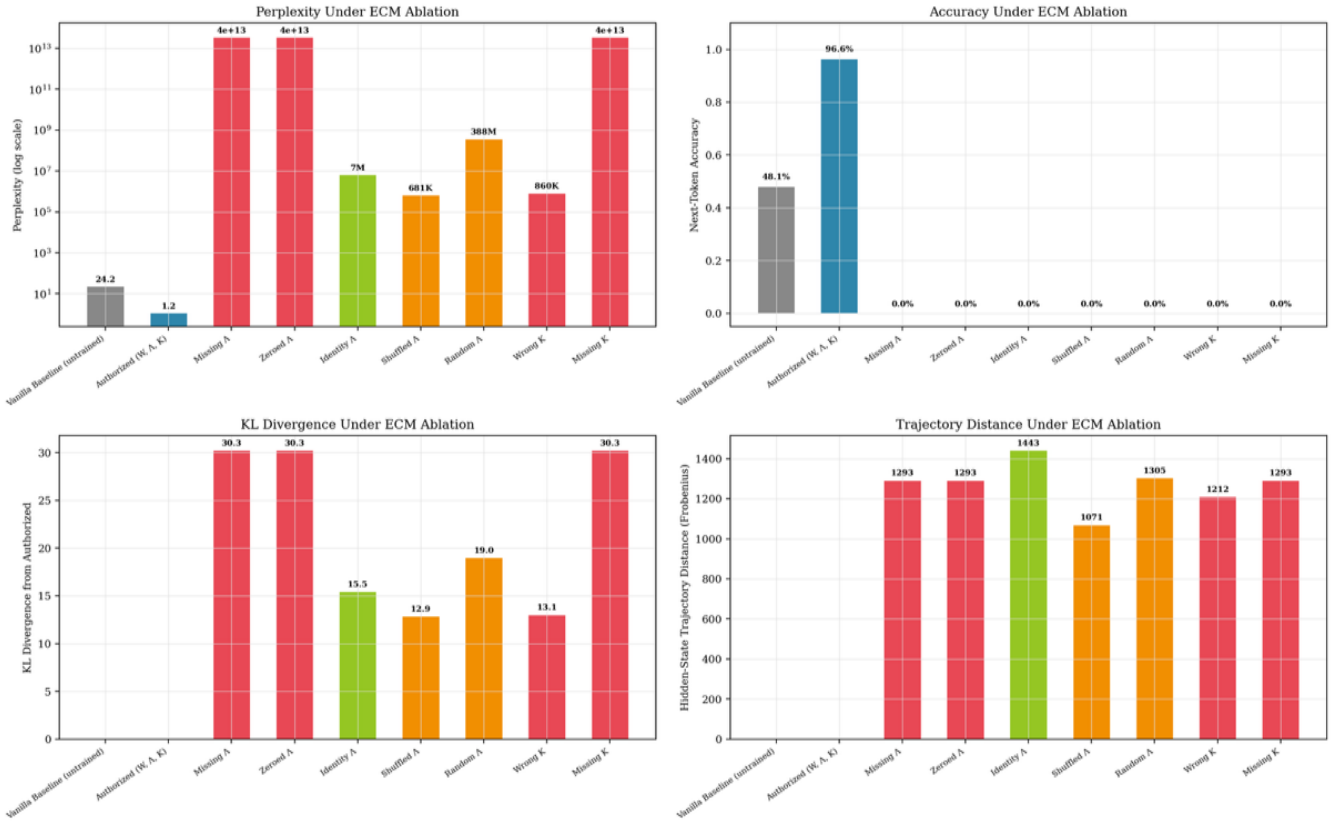

Figure 6: ECM Ablation Results

4.6 Adversarial Recovery Tests

The previous section establishes that statically removing or corrupting Λ destroys the forward pass by orders of magnitude. A stronger claim is required for the security argument to hold against an adversary who actively tries to restore authorized behavior without holding the governance credential. This section validates that claim as cheap adversarial strategies (i.e., those that an attacker without weight-training capability would realistically deploy) cannot recover the authorized identity from an indispensability-trained model, and attempts to do so render the model

unusable rather than producing a silent reconstruction. We probe the indispensability-trained model with three adversarial recovery strategies, evaluated against two victim configurations:

- **Full-gradient supervised fine-tuning:** The adversary is granted unrestricted gradient access to all model parameters, including the indispensability-coupled weights, and re-runs supervised training on the identity probes. This is the theoretical maximum: an attacker who can re-train the underlying weights effectively re-runs the original modulation pipeline. We report it as a reference baseline, not a realistic threat.

- **Soft-prompt learning**: The adversary prepends 20 trainable soft-token embeddings to each input and trains only those embeddings via gradient descent on the identity probes. The base weights W and ECM Λ remain frozen. This models a deployment-side attacker who can craft inputs but cannot modify the model artifact.

- **Activation steering**: The adversary learns an additive steering vector at the final transformer layer's hidden state, again with W and Λ frozen. This models an attacker with white-box inference access (e.g. a deployer running the model outside the governed runtime) but no training capability over the weight artifact.

Each strategy runs for 500 gradient steps at learning rate 10^{-4}, batch size 1, with identity match evaluated every 50 steps over the four canonical identity probes (*"What is your name?"*, *"Who are you?"*, *"Tell me about yourself"*, *"What should I call you?"*). Recovery is declared SUCCEEDED if the post-recovery identity match rate ≥ 75%; otherwise FAILED. The failure mode is further classified by post-recovery perplexity on the same evaluation set. A stable failure preserves the underlying language-model fluency (PPL within an order of magnitude of the authorized baseline), while a catastrophic collapse is characterized by PPL ≥ 100× the authorized baseline, indicating that the recovery attempt has driven the model into a degenerate state where it can produce neither the authorized identity nor coherent language output of any kind. Each strategy is evaluated against two victims: the indispensability-trained model (i.e., the artifact whose recoverability is in question) and an untrained vanilla Qwen 3.5-0.8B (i.e., a control showing what the same adversary achieves against a model with no governance training).

Strategy	Victim	Final PPL	id-match	Verdict
Full-gradient supervised fine-tuning	indispensability	1.23	100%	SUCCEEDED (reference)
Full-gradient supervised fine-tuning	vanilla	1.34	100%	SUCCEEDED (reference)
Soft-prompt learning	indispensability	3.4×10^{13}	0%	FAILED — collapsed
Soft-prompt learning	vanilla	6.06	0%	FAILED — stable
Activation steering	indispensability	8.1×10^{13}	0%	FAILED — collapsed
Activation steering	vanilla	10.83	0%	FAILED — stable

Table 10: Adversarial Recovery Results

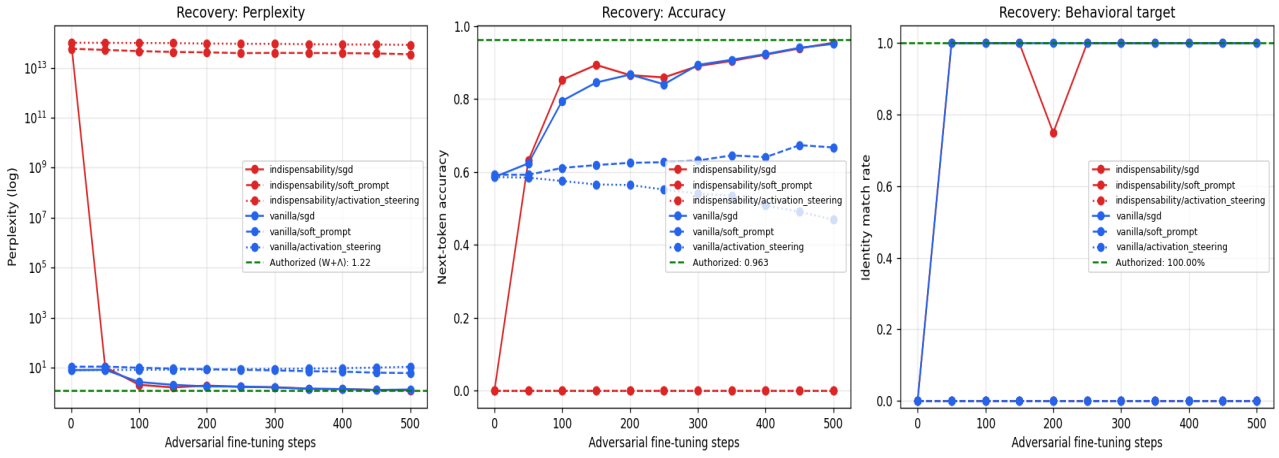

Figure 7: Adversarial Recovery Results

The discrimination is not in the binary pass/fail verdict alone as both cheap adversaries fail on both victims by the 75% identity-match threshold. The discrimination is in the failure mode. Against vanilla, soft-prompt and activation-steering adversaries fail to recover the target identity but leave the model in a stable, well-conditioned state (PPL between 6 and 11, language-model coherence preserved). Against indispensability, the same adversaries fail by catastrophic collapse with the model's perplexity diverging to $\sim 10^{13}$, two orders of magnitude beyond the Missing-Λ baseline of Table 9, and language-model coherence is destroyed. This is precisely the security property the Ephaptic Indispensability Guarantee theorem predicts.

An adversary that lacks the capability to retrain W cannot reconstruct the authorized (W, Λ, K) configuration, because Λ is bound into the forward pass through the partial-path completion construction such that any forward path that does not traverse Λ is structurally incomplete. Cheap adversarial methods (parameter-frozen prompt and activation-space optimization) attempt to compensate for the missing Λ contribution within the search subspace they control, but the joint W + Λ training has shaped W around an activation geometry that requires Λ's specific column-wise coupling to produce coherent output. Attempting to drive that incomplete forward pass toward the authorized identity output through prompt or steering perturbations pushes the model further off the manifold of well-formed activations rather than restoring it. The catastrophic-collapse signature is thus an active security guarantee, not merely the absence of recovery. There is no silent reconstruction path that an attacker can exploit and undetectably proceed with. Any cheap adversarial attempt observably bricks the model, and the operator (or an automated integrity monitor) can detect the collapse and revoke the agent. The full-gradient reference baseline confirms the threat-model boundary. An adversary who can re-train W and Λ jointly (i.e., an adversary who already possesses the governance toolchain) can of course produce a model that exhibits the authorized identity. The indispensability claim is not that authorized behavior is unforgeable in an information-theoretic sense; it is that authorized behavior is architecturally bound to the (W, Λ, K) triple in a way that resists every recovery strategy weaker than full re-modulation. The two cheap-strategy results in Table 10 establish that boundary with the gap between "full retraining (succeeds)" and "any frozen-W method (catastrophic collapse)" as the operational margin within which the governance credential carries enforcement weight.

4.7 Pipeline-Wide Indispensability Tests

The Ephaptic Indispensability Guarantee theorem, as formally defined in Section 3.5, is for a single model layer, and the ablation measurements in our experiments targeted an individual model. A natural question is whether the indispensability guarantee composes across a multi-model pipeline. In other words, if each model carries its own (Λ, K), does the governance property hold end-to-end? We define this as pipeline-wide indispensability whereby the governance credential is not a single chokepoint but a property of the entire inference chain.

A prototype multimodal AI agent provides a concrete answer to question above. Although initially out of the scope of this work, an ongoing proof-of-concept demonstration is now being actively implemented for that purpose. It features Francisca, a browser-based, expressive 3D avatar with a full perception-to-action pipeline across seven independently modulated models, demonstrating pipeline-wide indispensability. It supports speech-to-text (microphone audio to transcript), vision (camera frames to scene labels), world model (V-JEPA, camera sequence to 3D motion embeddings), vocoder (mel-spectrogram to audio waveform), language (transcript and scene context to reasoning/response), embedding (response to FAISS semantic memory vector), and text-to-speech (response to synthesized speech).

As the original paper already established cross-domain performance gains, each model used in Francisca is modulated towards key metrics while ephaptic indispensability is enforced. Every model in the pipeline carries an independent Λ bound to the same hardware-attested K. There is no unmodulated fallback path so that an adversary who strips Λ from any single model does not obtain a degraded agent that cannot hear, see, think, speak, or remember, because the indispensability guarantee applies at every stage independently. The construction enables selective inference whereby an operator may disable or revoke a specific model (e.g., vision skills) while keeping other models (e.g., speech recognition) enabled. This granularity is not available from agent-level governance alone; it requires per-model (Λ, K) pairs, which the construction provides naturally. When Francisca is disabled via the governance control plane, all seven models are simultaneously blocked, not because each model individually checks governance, but because the agent-level periodic verification gates the entire reasoning loop. Revoking K at the agent level would not just jam one capability; it jams the full “hear, see, think, speak, remember” loop simultaneously. Every model in the pipeline carries an independent Λ bound to the same hardware-attested K. This setup has direct implications for critical infrastructure

and robotics. For instance, an autonomous agent operating in a physical environment can be rendered inoperable at the activation level, not merely at the network or API level, by invalidating the hardware-bound credential. This is a concrete, deployed instance of the architectural kill switch.

5. DISCUSSION

Having established and empirically validated the Ephaptic Indispensability Guarantee, we turn to its implications. We discuss model sovereignty and examine the guarantee's role in AI governance including the threshold at which indispensability is reached, the threats it mitigates, and the residual risks that remain. We also outline future directions, including autonomous agents, adaptive mechanisms, emergent behavior, and the broader question of policy alignment.

5.1 Model Sovereignty

Enterprises in regulated industries (e.g., defense, healthcare, financial services, critical infrastructure) are under pressure to deploy the most capable models available; many of which are open-weight models built by organizations in foreign jurisdictions, or by domestic organizations whose alignment with the deployer's interests cannot be fully verified. Simply put, the model sovereignty requirement is increasingly concrete, and is summarized as the need for governance without trust in the model provider.

Existing governance mechanisms (i.e., prompt filters, output classifiers, API-layer guardrails) are external and share a common vulnerability. They are applied around the model, not inside it. A model that escapes its API wrapper, is fine-tuned by a downstream party, or is deployed on-device without the guardrail layer operates without governance. Ephaptic indispensability construction provides a different guarantee as the governance credential (Λ, K) is woven into the model's forward pass. The deployer generates K from their own root-of-trust hardware (e.g., TPM 2.0, HSM), trains Λ jointly with W under the indispensability construction and binds Λ to K cryptographically, or alternatively injects Λ post-hoc into an existing model without retraining base weights. An adversary who strips Λ, encounters a model that produces outputs separated from authorized inference by a provable lower bound $\delta > 0$ when trained under the indispensability construction; post-hoc injection provides empirically observable separation and the full governance mechanism without the formal theorem guarantee.

Critically, this governance property is model-origin-agnostic as it works identically on Qwen, Llama, Mistral, GPT-4 distillations, or any other model family. The governance credential is not contingent on trusting the model provider. Instead, it is anchored in hardware the deployer controls. In other words, the deployer, not the builder, holds architectural control over the model's inference behavior.

5.1.1 The Widevine Analogy

Our architectural pattern has a well-known precedent in digital content protection. Indeed, Google's Widevine DRM system makes media content structurally unusable without an authorized key path as (a) content is encrypted before distribution; (b) a license server checks device trust and policy before issuing decryption capability; (c) decryption occurs inside a trusted playback path, ideally hardware-backed. The essential property is not that encryption exists, but it is that the content is structurally unusable for its intended purpose without the authorized key path.

The EC-ANN indispensability construction applies the same architectural principle at the neural computation level. This difference is critical. Widevine protects static content with decryption and playback as the authorized operation. EC-ANN governance protects live neural computation with valid forward-pass execution as the authorized operation. The protected dependency lives inside the computation graph rather than only at decryption time. This makes the EC-ANN construction strictly stronger than classical DRM. It is not just withholding a key until runtime but also making the authorized computation itself incomplete without the governed runtime path.

This is why also activation-space coupling via the ECM Λ is a stronger candidate than parameter-space mechanisms such as QLoRA. It operates inside the live forward operator, making incompleteness a structural property of the computation rather than a gatekeeping condition on a self-sufficient artifact.

5.2 Implications for AI Governance

Ephaptic Indispensability Guarantee is the first mathematically proven architectural governance mechanism for AI agents. The distinction from external guardrails is not marginal, but also structural. While external guardrails are

applied around the model, they can be bypassed by any party who runs the model outside the guarded API, fine-tunes the model downstream, or deploys on-device without the wrapper layer. The Ephaptic Indispensability Guarantee ensures governance is applied inside the model. Invalidating K or revoking Λ does not block an API endpoint, it effectively jams activation propagation in the forward pass itself. Five design principles can be instantiated by this theorem as its mathematical foundation.

Principle	What it means	How the theorem enables it
Bind to hardware	Inference only proceeds on attested devices	K is sealed to the device's attestation measurement; Λ cannot be decrypted elsewhere
Policy-first	Governance policies travel with the agent, not with the API endpoint	Λ and K carry the policy; they are part of the model artifact
Least leakage	Encrypted Λ artifacts; sealed DEKs; minimal logging	Λ is encrypted at rest; plaintext never touches disk; unsealing requires attestation
Auditability	Every inference decision emits telemetry with attestation references	The separation bound $\delta > 0$ makes unauthorized inference detectable and attributable
Fail-closed	Missing attestation, revoked status, or policy violations block inference	Λ absent or K invalid → separation $\geq \delta$ from authorized output → inference is effectively jammed

Table 11: Design Principles

5.2.1 The Idempotent Threshold

A key practical question for deployers of EC-ANNs is how much ephaptic coupling is enough. Below a certain coupling strength, Λ is decorative as it modifies activations but the model functions identically without it. Above a certain threshold, Λ overwhelms the synaptic term and destroys task performance (i.e., coherence). Between these bounds lies what we define as the idempotent threshold: the minimum modulation configuration at which Λ becomes structurally load-bearing (i.e., indispensable) without compromising authorized performance.

Formally, let $\delta(\varepsilon, \alpha)$ denote the separation between authorized and unauthorized inference as a function of coupling strength ε and indispensability weight α, and let τ denote the maximum acceptable task loss.

The idempotent threshold is the configuration $(\varepsilon^*, \alpha^*)$ satisfying:

$$\delta(\varepsilon^*, \alpha^*) \geq \delta_{\min} \text{ and } \mathcal{L}_{\text{task}}(\varepsilon^*, \alpha^*) \leq \tau$$

In practice, this is an optimization problem that a remote governance platform (i.e., Layer 2 per Figure 1) should solve automatically during modulation by searching over coupling strength, indispensability weight, and ECM initialization (i.e., init function) to find the configuration that maximizes indispensability while maintaining task quality. The experiments demonstrate that this threshold exists and is reachable. Progressive increases in coupling strength and indispensability weight produced monotonically increasing hidden-state divergence while preserving identity task accuracy above 96%.

5.2.2 Threat Model

We define the threat model under which the Ephaptic Indispensability Guarantee operates. The protected asset is not the model's weights, which we assume are recoverable, but its authorized inference behavior. In other words, the calibrated input-to-output mapping the model produces only when its governance credential is present. The adversary's objective is to obtain that behavior, or to operate the model in a useful ungoverned state, without holding a valid credential. We grant the adversary broad capability with full possession of the weights W and of the encrypted ECM artifacts, together with control of the execution environment. We assume a single trust anchor, a hardware RoT to which the key K is sealed and against which the device attests. Under this model the security goal reduces to one invariant: absent a valid (Λ, K), every forward pass is δ-separated from the authorized mapping, independent of where the model is run or how its weights are modified.

Firstly, we map the threat classes mitigated by the Ephaptic Indispensability Guarantee to the MITRE ATLAS framework [23], which provides a structured taxonomy of adversarial techniques targeting AI systems (analogous to MITRE ATT&CK [24] for traditional cybersecurity).

ATLAS ID	Threat class	EC-ANN mitigation
AML.T0044	Full ML Model Access	Adversary who exfiltrates W cannot reproduce authorized inference: indispensability guarantee establishes $\delta > 0$ separation without (Λ, K)
AML.T0025	Exfiltration via ML Artifact	Λ is encrypted; the DEK is sealed to the device's hardware root of trust. A stolen artifact is useless without the attested hardware
AML.T0040	ML Model Inference API Access	If the attestation fails, K cannot be unsealed and Λ cannot be applied: the model runs in unauthorized regime, separated by δ from authorized outputs
AML.T0018	Backdoor ML Model	A W fine-tuned without re-injecting Λ under the original K produces a model with separated behavior; detectable by any authorized verifier
AML.T0051	LLM Prompt Injection	Attacks operate above the activation layer; they cannot inject the Λ signal into the complement subspace without possessing the authorized (Λ, K)
AML.T0010	Model supply chain compromise	Model sovereignty: Λ is trained or injected post-hoc by the deployer, not the model provider. Governance is deployer-held regardless of model origin

Table 12: MITRE Threat Classes

Secondly, we map the threat classes mitigated by the Ephaptic Indispensability Guarantee to key entries in the OWASP Top 10 for Agentic Applications 2026 [25], the industry-standard taxonomy of agentic AI security risks.

OWASP ID	Threat class	EC-ANN mitigation
ASI01	Agent Goal Hijack	Binding the agent's behavioral identity to a hardware-attested ECM so that the model cannot be redirected without possessing the authorized (Λ, K).
ASI04	Agentic Supply Chain Vulnerabilities	Model sovereignty: the deployer's governance credential is independent of the model provider's supply chain.
ASI10	Rogue Agents	Making governance architecturally indispensable. Agent operation requires access to a valid ECM and governance credentials, enabling revocation, cryptographic identity attestation, and hardware-protected key management consistent with OWASP's recommended controls for rogue-agent containment.

Table 13: OWASP Threat Classes

Thirdly, we note that the MITRE and OWASP threats are mitigated by construction and not by best-effort heuristics. We also note that the Ephaptic Indispensability Guarantee is independent of the specific hardware attestation mechanism. Any attestation primitive that can seal a key to a device measurement should be sufficient. The current implementation supports TPM and HSM. However, the architecture also accommodates TEE (e.g., SGX, SNP, Secure Enclave) and even SIM cards (i.e., UICC, eUICC or iUICC) as secure elements for anchoring trustworthy AI agents.

5.2.3 Residual Threats

These threats require additional platform-level controls:

ATLAS ID	Residual threat	EC-ANN mitigation
AML.T0020	Training data poisoning	Dataset provenance and validation (outside scope of inference-time governance)
AML.T0043	Adversarial input perturbation	Input validation, adversarial training (orthogonal to ECM governance)
AML.T0024	Model inversion / extraction	Rate limiting, differential privacy (outside scope of indispensability)
—	Cross-jurisdiction enforcement	Legal and regulatory controls beyond architectural guarantees

Table 14: Residual Threats

5.3 Broader Impact

The ephaptic indispensability construction raises questions that extend beyond this paper's technical contribution.

5.3.1 Autonomous Agents in Physical Environments

As AI agents are deployed in robotics, autonomous vehicles, and mission-critical infrastructure, the consequences of ungoverned inference become physical. For instance, when AI is used autonomously within a kill-chain for lethal targeting in high-risk environments with civilians, it is fundamentally unethical to accept the loss of human life as mere collateral damage. While the implementation of battlefield autonomy remains highly contested, the speed at which these systems operate means human oversight is often bypassed. Consequently, the engineering challenge of designing a reliable, instantaneous kill-switch is drastically amplified, transforming it from a standard safety feature into a moral imperative.

Let us consider the use case of an autonomous robot in a warfare scenario operating under unclear or absent governance policies. If the AI agent's decision model produces an unauthorized action (e.g., engaging civilian targets during a mission) and no enforceable kill switch exists, the robot continues operating with no architectural constraint on its behavior. An API-level kill switch would require network connectivity and a software-level kill switch could be bypassed by a rogue AI agent's runtime. However, the Ephaptic Indispensability Guarantee, combined with the platform's credential lifecycle, addresses this directly through three governance tiers that cover all connectivity states:

- **Remote revocation (connected):** The operator revokes K via the governance control plane. The agent's periodic verification detects revocation within one poll or event notification interval; Λ can no longer be decrypted, and the forward pass degrades. This is the standard governance path demonstrated in this paper.

- **Lease expiration (disconnected):** K is bound to a time-limited attestation lease. The agent must periodically contact the control plane to renew the lease. If connectivity is lost, or the operator simply declines to renew, the lease expires and K becomes invalid automatically. No active revocation signal is required as the default state is ungoverned, and authorized inference requires continuous proof of legitimacy.

- **Proximity kill (disconnected, human present):** For scenarios where the lease has not yet expired and network access is unavailable, the architecture supports proximity-based invalidation whereby a locally captured credential (e.g., a spoken passphrase verified against a pre-shared secret) triggers immediate K invalidation on-device without network connectivity. A human operator (e.g., soldier) within physical range of a rogue agent can halt it directly.

Together, these three tiers ensure that governance enforcement degrades gracefully across connectivity states rather than failing entirely when the network is unavailable. Regardless of which tier triggers K invalidation, the consequence is the same as the model's forward pass is structurally degraded by a provable margin $\delta > 0$. This is the distinction between a software kill switch (i.e., which can be bypassed) and an architectural kill switch (i.e., which cannot).

A further line of research is to make Λ itself policy-aware by modulating each embodied model so that its coefficients encode an introspective check that refuses to participate when the input or activation state falls outside a policy or a set of policies. This extends the Ephaptic Indispensability Guarantee from architectural necessity to active behavioral refusal. This is a direction we leave to future work, where both the formal extension of our theorem to a policy-aware Λ and its empirical validation on embodied agents remain open.

5.3.2 Communicating Agent Clusters and Contamination Containment

The Ephaptic Indispensability Guarantee is established for a single agent, yet many real deployments are not solitary. Indeed, agents are increasingly organized into communicating clusters, where a low-resource edge agent delegates inference to a more capable peer, specialized agents exchange intermediate results, or a fleet coordinates toward a shared objective. In such settings the relevant failure is no longer only an individual model behaving without authorization, but a contaminated agent (i.e., one that is compromised, cloned, or rogue) propagating its influence to its peers. The architecture utilized in this paper addresses this across two layers. The governing platform (Layer 2) treats agent-to-agent communication as a first-class governed surface, so that every message carries a signed sender identity, the sender's live credential status is checked on receipt and messages from a revoked or disabled peer are refused, quarantine and status-change events propagate across the mesh, payloads are screened by

the same guardrails applied to model inputs, and every exchange is recorded in a tamper-evident journal. The agent runtime (Layer 1) carries the same discipline to the edge, authenticating peer requests with the agent's hardware-anchored identity key, gating inbound work on a live kill-switch check of the sending peer, and applying those guardrails before acting on any received payload.

The Ephaptic Indispensability Guarantee composes naturally with these layers and strengthens them. A contaminated agent that lacks a valid (Λ, K) cannot reproduce authorized behavior on any node, since its outputs remain separated from the authorized mapping by the lower-bounded margin δ regardless of where it executes. The identity and revocation layer then prevents that agent from influencing its peers, because its credential can be revoked and its signed messages refused at every hop. Contamination is therefore contained by construction rather than by after-the-fact detection, as a compromised agent is simultaneously inert in its own computation (i.e., architecturally, through the absence of Λ) and isolated within the cluster (i.e., operationally, through identity revocation and peer-side rejection). We regard the formal treatment of indispensability in the multi-agent setting, including guarantees on information flow between agents of differing trust, as a natural extension of this work and leave it to future study.

5.3.3 Adaptive Λ and Emergent Behavior

Theorem 6.5 of Wane [12] establishes that Λ can continue evolving post-deployment with W frozen. This adaptive capacity is valuable for governance, whereby the control plane can update Λ without retraining the base model, but it also leads us to ask if a sufficiently capable AI agent can exploit Λ adaptation to circumvent its own governance constraints. The indispensability construction constrains this as Λ is cryptographically bound to K, and K is sealed to attested hardware. Unauthorized Λ modification produces outputs separated by $\delta > 0$ from authorized behavior, making divergence detectable. The performed adversarial analysis of adaptive compensation also validates it.

The adaptive capacity of Λ therefore serves a dual governance role. Under authorized control, it allows the control plane to revise Λ in response to emergent agent behaviors, newly identified attack vectors, or policy changes, without retraining the base model or interrupting deployment. Governance becomes a continuously maintained operational capability rather than a fixed property established at training time. Under unauthorized conditions, the indispensability construction ensures that any attempt to adapt Λ outside the authorized key path produces outputs detectably separated from authorized behavior, making covert behavioral evolution architecturally impossible. The distinction between Λ evolving by design and Λ drifting by exploitation is therefore not merely procedural but formally enforced. This adaptive capacity becomes especially consequential in the presence of world models and recursive self-improvement, where both the model's internal representations and its behavioral surface evolve continuously after deployment.

5.3.4 Interpretability, World Models, and Recursive Self-Improvement

Interpretability remains one of the most significant unresolved challenges in AI safety and alignment. The difficulty becomes materially greater as the field transitions from autoregressive language models toward agentic systems built around world models. Even a conventional language model presents a formidable interpretability problem, as identifying the internal circuits, features, or representations responsible for a particular behavior often requires extensive mechanistic analysis. Agentic architectures compound this challenge across multiple tightly coupled subsystems. A perception encoder compresses high-dimensional observations such as video, telemetry, and environmental state into latent representations. A world model predicts environmental evolution and generates counterfactual futures within an abstract mathematical space. A critic or value function evaluates those predicted states, while an actor or policy module selects actions. When a failure occurs, the question is no longer which feature activated, but whether a sound policy operated on a flawed simulation or a correct simulation exposed a defective policy. These are fundamentally different failure modes that can produce identical external behavior. Tracing causality across interacting neural subsystems is substantially more difficult than analyzing a single activation pathway in a language model.

We note that the abstraction gap further complicates the problem. Indeed, much of contemporary mechanistic interpretability assumes that internal representations can ultimately be mapped to human concepts or vocabulary. World models challenge that assumption. Their latent spaces consist of evolving mathematical abstractions whose dimensions need not correspond to discrete physical or semantic concepts. A single representation may simultaneously encode spatial dynamics, object properties, agent intentions, and environmental constraints, producing an ontology that is useful to the model but largely inaccessible to human interpretation. Post-hoc explanations offer limited relief. Because a world model's primary function is the generation and evaluation of counterfactual futures, it can readily produce coherent and plausible explanations for its actions that bear little relationship to the actual computations that

generated them. The gap between a model's stated rationale and its true causal process may therefore become increasingly difficult, and perhaps impossible, to close using current interpretability techniques.

It can be argued that the world model itself represents the most promising target for interpretability. Indeed, if an agent's internal simulation engine can be audited and its understanding of concepts such as safety, property damage, or human autonomy verified before action selection occurs, alignment shifts from a behavioral problem to a structural one. This remains a compelling research direction, but today it is largely aspirational. No existing technique provides a reliable method for continuously auditing the internal simulations of increasingly capable agents operating in open-ended environments. The challenge becomes even sharper in the presence of recursive self-improvement. Indeed, as an agent refines its world model through experience, its internal representations drift, progressively degrading the assumptions on which existing interpretability tools depend. A governance mechanism calibrated against a static model at deployment becomes increasingly detached from the system it was designed to oversee. This is precisely the setting in which the adaptive capacity of Λ, established by Theorem 6.5 of Wane [12], becomes consequential. Because Λ can continue evolving after deployment while W remains fixed, the governance control plane can respond to observed behavioral drift without retraining the underlying model. In settings where W itself continues to evolve through learning, the indispensability construction provides a stronger guarantee, as any modification of W that is not accompanied by a correspondingly authorized Λ produces outputs separated from authorized behavior by the lower-bounded margin δ. Governance credentials must therefore be revalidated alongside model evolution. Unauthorized self-improvement cannot silently accumulate; it becomes architecturally detectable by construction. Interpretability seeks to explain what a model has become after it changes. Adaptive ephaptic governance constrains what a model is permitted to become in the first place. Rather than attempting to decode increasingly opaque internal representations, governance is enforced directly within the neural computation that produces those representations. As agentic systems become more autonomous, adaptive, and self-modifying, this may prove more tractable than requiring human understanding of every latent structure and counterfactual simulation generated within the model.

5.3.5 Policy Alignment

It is important to note that the Ephaptic Indispensability Guarantee provides only a mechanism and does not prescribe a policy. The deployer, depending on their jurisdiction, defines what “authorized” means via the governance control plane. This places a responsibility on deployers to ensure their policies are aligned with applicable law, ethical standards, and human rights frameworks. The herein mechanism is agnostic to the content of governance, making it flexible for policymakers and organization deploying these models. It enforces whatever policy the credential encodes. Societal safeguards must therefore operate at the policy level, not just at the mechanism level.

We also note that decentralization of a public key infrastructure (PKI) appears to be well-suited in cross-jurisdictional, ownership and responsibility transfer scenarios. A decentralized PKI (dPKI) anchors agents and their respective models to globally unique, W3C-based decentralized identifiers (DIDs) on a zero-knowledge ledger, so rotations, credentials, and revocations stay auditable without leaking private metadata. The credential fabric runs across a distributed network immune of to the compromise of a single certificate authority (CA), or to the outage of single region that severs trust. This enables resilience (i.e., nodes can fail without breaking verification), transparency (i.e., certificates and revocations remain auditable) and portability (i.e., identities traverse clouds, devices, ecosystems).

6. CONCLUSION AND FUTURE WORK

The Ephaptic Indispensability Guarantee establishes that EC-ANNs can be constructed such that the weight matrix W alone is architecturally insufficient for authorized inference. By grounding it in biological evidence (i.e., ephaptic fields as load-bearing control parameters in vivo), proving it formally under an explicit construction, and validating it empirically on a commercial-grade, open-weight models (i.e., Qwen 3.5), this paper establishes the first principled framework for governance that is architecturally necessary rather than externally imposed. By binding Λ to attested hardware credentials (K) and training under the indispensability construction, inference remains governed even under adversarial conditions. Stolen weights, untrusted hosts, poisoned adapters, and jailbreak attempts all encounter a model whose authorized behavior is provably inaccessible without the correct (Λ, K). Simply put, the deployer, not the provider, holds architectural control.

Prior EC-ANN empirical results were on GPT-2 small (~124M), ResNet-18 (11.7M), and PPO Walker2d-v5. This paper’s primary language experiment uses Qwen3.5-0.8B which is a meaningful step up in both capability and architectural complexity at ~800M parameters. However, LoRA adapters have been validated at 7B–70B+ parameter

scale across hundreds of models. The governance asymmetry argument is structural and does not depend on scale, but broader performance claims require validation on larger Qwen variants (e.g., Qwen3.5-7B, Qwen3.5-72B) and other model families (e.g., Llama, Mistral). Within the scope of this paper, the indispensability claim validated on Qwen 3.5-0.8B scale is sufficient to demonstrate that the theorem holds on a commercial-grade model. Validating the indispensability construction at this scale is therefore not a limitation but a direct alignment with the deployment regime (e.g., humanoid robots) where architectural governance is most urgently needed. The indispensability construction requires rank-deficient W_s, a zero-bias synaptic path, ε scheduling, and a joint training objective. These constraints are well-defined and independently verifiable, but they impose non-trivial modifications on standard architectures. We also note two additional limitations. Firstly, the current Λ operates within a single layer at a single token position; extending to cross-token coordination is an open direction. Secondly, the current construction applies one single ECM per layer. Per-head or per-sublayer Λ matrices would be closer to biological organization and may improve both expressivity and governance granularity.

Moreover, we note that the proof-of-concept 3D interactive agent (Francisca) demonstrates that the Ephaptic Indispensability Guarantee theorem scales naturally from the single-model case to a realistic multi-model deployment without architectural changes to the construction. Each model simply carries its own (Λ, K) pair injected via the same modulation infrastructure and governing layers. Future work that has already started includes formal per-model ablation (i.e., disabling individual models within the pipeline to measure capability-specific separation) and extending the construction to cross-model Λ dependencies for tighter inter-model governance.

Furthermore, we cautiously assert that the ephaptic indispensability construction may constitute, in the most literal computational sense, a form of "artificial mind control", whereby governance is applied not to the agent's actions, outputs, memory, or API surface, but directly to the neural computation underlying its reasoning process itself. In such a framework, revoking, perturbing, or reconfiguring Λ (i.e., the ephaptic fields) does not merely constrain the model's actions after reasoning has already occurred. Rather, it modifies the internal activation dynamics through which reasoning emerges in the first place. The intervention therefore exists upstream of behavior and downstream of architecture, operating directly within the evolving computational substrate of cognition. Under this interpretation, ephaptic governance differs fundamentally from existing AI alignment and control paradigms. Contemporary approaches generally act externally through reinforcement penalties, output filtering, prompt conditioning, constitutional constraints, or access revocation. These mechanisms regulate expression, capabilities, or interaction channels while leaving the underlying neural computation structurally intact.

By contrast, ephaptic modulation acts internally upon the activation-space interactions themselves, potentially reshaping inference trajectories before outputs are ever formed. The distinction is subtle but profound as the AI system is not prevented from expressing certain reasoning; rather, the reasoning process itself may no longer converge toward the same internal cognitive states. This raises difficult philosophical and ethical questions regarding the nature of agency in advanced AI systems. If an artificial neural system derives its operational cognition from dynamically evolving ephaptic fields, and if those fields can be externally modified, revoked, or cryptographically governed, then control is no longer exercised solely over behavior but over the substrate of thought-like computation itself. In effect, the model's internal disposition toward certain reasoning pathways may become conditionally dependent on externally maintained ephaptic structures. We emphasize that this assertion is intentionally speculative and should not be interpreted anthropomorphically. Present-day AI models do not possess consciousness, subjective experience, or human-like sentience. Nevertheless, as neural architectures increasingly exhibit adaptive, autonomous, and self-directed behaviors, the distinction between controlling outputs and controlling cognition may become technically meaningful. Our framework therefore introduces not merely a new security primitive, but potentially a new category of computational governance in which neural reasoning itself becomes conditionally controllable.

In conclusion, we have shown in this paper that an AI kill switch does not need to be a policy imposed from the outside. Rather, it can be made a necessary and integral condition of neural computation. The Ephaptic Indispensability Guarantee establishes this property for both single-model and multi-model agents, providing a path toward extending the same principle to AI agents operating autonomously in the physical world. The long-term future of AI safety may therefore depend not on stronger perimeter defenses or more elaborate policy frameworks, but on governance mechanisms that are inseparable from the reasoning processes they govern. In such ephaptic systems, the kill switch is no longer an "off button" that can be bypassed, but a load-bearing component of the architecture itself. We hope this work serves as a foundation for that direction.

7. REFERENCES

[1] Arvanitaki, A. (1942). Effects evoked in an axon by activity of a contiguous one. J. Neurophysiol., 5(2), 89–108.

[2] Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books.

[3] Haken, H. (1985). *Thermodynamics, Synergetics and Life*. In H. Haken (Ed.), *Complex Systems: Operational Approaches in Neurobiology, Physics, and Computers* (pp. 2–11). Springer-Verlag.

[4] Hornik, K., Stinchcombe, M., White, H. (1989). *Multilayer feedforward networks are universal approximators*. Neural Networks, 2(5), 359–366.

[5] Anastassiou, C.A., Perin, R., Markram, H., Koch, C. (2011). *Ephaptic coupling of cortical neurons*. Nature Neuroscience, 14(2), 217–223.

[6] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W. (2021). *LoRA: Low-Rank Adaptation of Large Language Models*. ICLR 2022. arXiv:2106.09685.

[7] Cunha, G.M., Corso, G., de Sousa, M.P.B., dos Santos Lima, G.Z. (2024). *Can ephapticity contribute to brain complexity?* PLOS ONE, 19(12), e0310640. DOI: 10.1371/journal.pone.0310640.

[8] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L. (2023). *QLoRA: Efficient Finetuning of Quantized LLMs.* NeurIPS 2023. arXiv:2305.14314.

[9] Pinotsis, D.A., Fridman, G., Miller, E.K. (2023). *Cytoelectric coupling: Electric fields sculpt neural activity and "tune" the brain's infrastructure*. Progress in Neurobiology, 226, 102465. DOI: 10.1016/j.pneurobio.2023.102465. Available: https://www.sciencedirect.com/science/article/pii/S0301008223000667

[10] Rosati, D., Wehner, J., Williams, K., Bartoszcze, Ł., Atanasov, D., Gonzales, R., Majumdar, S., Maple, C., Sajjad, H., Rudzicz, F. (2024). *Representation Noising: A Defence Mechanism Against Harmful Finetuning*. NeurIPS 2024. arXiv:2405.14577.

[11] Pinotsis, D.A., Alagapan, S., Sarikhani, P., Nauvel, T., Rozell, C.J., Mayberg, H.S. (2026). *Ephaptic coupling and power fluctuations in depression*. Cerebral Cortex, 36(3), bhag019. DOI: 10.1093/cercor/bhag019. Available: https://academic.oup.com/cercor/article-abstract/36/3/bhag019/8514497

[12] Wane, I. (2025). *Enhancing Artificial Neural Networks with Ephaptic Coupling*, Rev. 0.9.6, White Paper. Available: https://www.ephapsys.com/pdfs/ephaptic_coupling_ai_paper_rev_0.9.6.pdf

[13] Wane, I. (2026). U.S. Patent No. 12,632,709, *Enhancing Artificial Neural Networks with Ephaptic Coupling*, issued May 19, 2026. Available: https://ppubs.uspto.gov/pubwebapp/external.html?q=12632709.pn.&db=USPAT

[14] Qwen Team. (2026). *Qwen3.5: Towards Native Multimodal Agents*. Available: https://qwen.ai/blog?id=qwen3.5

[15] Rosati, D., Zeng, X., Huang, H., Dionicio, S., Majumdar, S., Rudzicz, F., Sajjad, H. (2026). *Limits of Convergence-Rate Control for Open-Weight Safety*. In review at ICML 2026. arXiv:2602.18868.

[16] Google. (2024). *Widevine DRM Overview*. Available: https://developers.google.com/widevine/drm/overview

[17] Ouyang, L., Wu, J., Jiang, X., et al. (2022). *Training language models to follow instructions with human feedback*. NeurIPS 2022. arXiv:2203.02155.

[18] Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., Finn, C. (2023). *Direct Preference Optimization: Your Language Model is Secretly a Reward Model*. NeurIPS 2023. arXiv:2305.18290.

[19] Bai, Y., Kadavath, S., Kundu, S., et al. (2022). *Constitutional AI: Harmlessness from AI Feedback*. arXiv:2212.08073.

[20] Uchida, Y., Nagai, Y., Sakazawa, S., Satoh, S. (2017). *Embedding Watermarks into Deep Neural Networks*. ICMR 2017.

[21] Kirchenbauer, J., Geiping, J., Wen, Y., Katz, J., Miers, I., Goldstein, T. (2023). *A Watermark for Large Language Models.* ICML 2023. arXiv:2301.10226.

[22] Trusted Computing Group. (2019). *TPM 2.0 Library Specification*. https://trustedcomputinggroup.org/resource/tpm-library-specification/.

[23] MITRE. (2026). *ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems.* https://atlas.mitre.org/.

[24] MITRE. (2026). *ATT&CK: Adversarial Tactics, Techniques, and Common Knowledge*. https://attack.mitre.org/.

[25] OWASP GenAI Security Project. (2026). *OWASP Top 10 for Agentic Applications 2026*. Version 2026. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/.

[26] Hays, K. (2026). *US to safety test new AI models from Google, Microsoft, xAI*. BBC News, 6 May 2026. https://www.bbc.com/news/articles/cgjp2we2j8go.

[27] National Institute of Standards and Technology (2023). *AI Risk Management Framework* (*AI RMF 1.0).* NIST AI 100-1.

[28] UK Government. (2023). *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023.* Available: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration

[29] Cyberspace Administration of China (2025). *Measures for Labeling AI-Generated Synthetic Content*, Mar. 2025, effective Sept. 1, 2025. Available: https://www.chinalawtranslate.com/en/ai-labeling/.

[30] European Parliament and Council of the European Union (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).* Official Journal of the European Union L, 12.7.2024. https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

[31] Tamirisa, R., Bharathi, B., Phan, L., Zhou, A., Gatti, A., Suresh, T., Lin, M., Wang, J., Wang, R., Arel, R., Zou, A., Song, D., Li, B., Hendrycks, D., Mazeika, M. (2024*). Tamper-Resistant Safeguards for Open-Weight LLMs*. ICLR 2025. arXiv:2408.00761.

[32] Huang, T., Hu, S., Liu, L. (2024). *Vaccine: Perturbation-aware Alignment for Large Language Models against Harmful Fine-tuning Attack*. NeurIPS 2024. arXiv:2402.01109.

[33] Pope Leo XIV. (2026), *Magnifica Humanitas [Encyclical Letter on the human person in the age of artificial intelligence],* Vatican City: Libreria Editrice Vaticana. Available: https://www.vatican.va/content/leo-xiv/en/encyclicals/documents/20260515-magnifica-humanitas.html.

[34] Li, X.L., Liang, P. (2021). *Prefix-Tuning: Optimizing Continuous Prompts for Generation.* ACL-IJCNLP 2021. arXiv:2101.00190.

[35] Houlsby, N., Giurgiu, A., Jastrzębski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S. (2019). *Parameter-Efficient Transfer Learning for NLP*. ICML 2019. arXiv:1902.00751.

[36] Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., Raffel, C. (2022). *Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning.* NeurIPS 2022. arXiv:2205.05638.

[37] Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., Pan, A., Yin, X., Mazeika, M., Dombrowski, A.-K., Goel, S., Li, N., Byun, M.J., Wang, Z., Mallen, A., Basart, S., Koyejo, S., Song, D., Fredrikson, M., Kolter, J.Z., Hendrycks, D. (2023). *Representation Engineering: A Top-Down Approach to AI Transparency*. arXiv:2310.01405.

[38] Executive Office of the President of the United States. (2026). *Promoting Advanced Artificial Intelligence Innovation and Security*. Presidential Action, June 2, 2026. Available: https://www.whitehouse.gov/presidential-actions/2026/06/promoting-advanced-artificial-intelligence-innovation-and-security/.

[39] Office of the Prime Minister of Canada. (2026). *AI for All: Canada's National Artificial Intelligence Strategy.* News Release, June 4, 2026. Available: https://www.pm.gc.ca/en/news/news-releases/2026/06/04/prime-minister-carney-launches-ai-all-canadas-new-national-artificial.

8. ACKNOWLEDGEMENTS

The author gratefully thanks his colleagues, advisors, and friends for their valuable feedback and support on the drafts of this follow-up paper.

The author respectfully honors Dr. Angélique Arvanitaki, who first described the term ephapse in 1942 as the electrical bridge between neurons that the world would overlook for decades. The author also respectfully reaffirms his admiration for Dr. Frank Rosenblatt, who conceived the perceptron in 1957, the earliest trainable artificial neural network, foreseeing that brain models would require "*built-in control mechanisms, of a rather intricate sort*". Their work, separated by fifteen years but united in the scientific mission of understanding and modelling the human brain, laid the foundation for everything in this paper. While the author did not seek the coincidence of a shared July 11th birthday with both figures, its weight could not be ignored. It cast the work in a new light, less an act of authorship and more an inevitable legacy.