

Attractor Calculus: A Formal Operator Language for Steering Cognition Across Substrates

Attractor Calculus: A Formal Operator Language for Steering Cognition Across Substrates

Abstract

Prompt engineering, neural stimulation, and cognitive intervention all face the same underlying problem: steering an information-processing system into a desired dynamical state, with no principled language for saying what "steering" means or what it should cost. Practitioners in each domain currently work system-by-system, with no way to state whether a control primitive discovered in one substrate transfers to another. This paper formalizes that gap. We define an Attractor Calculus: a small operator set (`CREATEATTRACTOR`, `DEEPENBASIN`, `SHIFTBOUNDARY`, `PERTURB`, `HOLD`) over attractor-based dynamical systems, together with a cost functional and a composition algebra. We derive closed-form minimal-energy costs for basin transitions in a canonical bistable potential $V(x) = x^4/4 - x^2/2$: an impulse crossing costs $E_{\text{imp}} = 2\Delta V = 0.5$, a constant-tilt crossing costs at the rate $u_c^2 = 4/27 \approx 0.148148$ per unit time, and the break-even horizon between the two strategies is $T^* = 27/8 = 3.375$. Under noise intensity $D = 0.05$ and a dwell-time reliability requirement of failure probability $\leq 10^{-3}$ over $T_d = 10^6$ time units, `DEEPENBASIN` must add $\delta d \approx 0.786138$ of barrier depth, raising impulse cost by $2\delta d \approx 1.572277$. We prove a partial universality result: impulse-operator cost depends only on the invariant barrier height ΔV , while tilt-operator cost depends on substrate geometry, so the operator set splits into portable and substrate-specific classes. This gives practitioners a testable criterion for which control levers should transfer across substrates and which must be re-derived per system.

1. Introduction

Anyone who deploys large language models, designs brain-stimulation protocols, or runs behavioral interventions faces the same operational problem: a target system sits in some dynamical regime, the desired behavior corresponds to a different regime, and the practitioner must find an input sequence that moves the system between the two reliably and cheaply. Today this is done by hand. Prompt engineers iterate on wording until outputs improve, with no vocabulary for stating what a prompt did to the model's internal state or why a fix that worked on one model fails on another. Neuromodulation and neurofeedback protocols are tuned per patient and per device. Behavioral interventions are designed case by case. In each domain the absence of a common control language blocks three concrete capabilities: predicting the cost of a steering maneuver before attempting it, reusing a control primitive discovered in one system inside another, and auditing whether a stimulus sequence is doing semantic work or merely pushing latent state.

The research idea this paper develops conjectures that information-processing systems, whether large language models or brains, are attractor-based dynamical systems, and that a formal operator calculus over attractor basins could compile high-level cognitive goals into minimal-energy stimulus sequences. If the conjecture holds, a control primitive validated in a toy recurrent network should transfer, with predictable cost rescaling, to a transformer or a neural circuit, and the question "is there a substrate-independent machine code of cognition?" becomes an empirical program rather than a slogan. If it fails, the failure itself is informative: it would show that steering levers are irreducibly substrate-specific and that cross-substrate transfer claims in prompting and neurostimulation should be treated with suspicion.

This paper does not settle the conjecture empirically. It builds the formal layer that makes the conjecture testable, and it derives the first closed-form results that the formal layer predicts. The contributions are:

- A definition of the Attractor Calculus: a state-space semantics, five operators (*CREATEATTRACTOR*, *DEEPENBASIN*, *SHIFT_BOUNDARY*, *PERTURB*, *HOLD*), a quadratic cost functional, and a composition algebra with an additivity theorem (Section 3).
- Closed-form minimal-energy derivations for basin transition in a canonical bistable potential: impulse cost $E_{\text{imp}} = 2\Delta V$, tilt cost rate $u_c^2 = 4/27$, break-even horizon $T^* = 27/8$, and their generalization to the family $V(x) = x^4/4 - ax^2/2$ giving $T^*(a) = 27/(8a)$ (Section 4).
- A noise-robustness derivation: under noise intensity D and a failure-probability requirement ε over dwell time T_d , the required basin depth is $\Delta V' \geq D \ln(T_d/(-\ln(1 - \varepsilon)))$, with the numerical consequence $\delta d \approx 0.786138$ for the assumed operating point (Section 4).
- A universality criterion and a partial classification of the operator set: *PERTURB* cost is a function of the invariant ΔV alone and is therefore portable across substrates in the defined sense, while *SHIFT_BOUNDARY* cost depends on local barrier geometry and is substrate-specific (Sections 4 and 5).
- An engineering pathway: the smallest experiment that would de-risk or falsify the calculus on a real trained network, the readiness level of the current work, and the binding constraints (Section 7).

2. Background and Related Work

The calculus proposed here sits at the intersection of three literatures: dynamical systems theory applied across substrates, formal language methods applied to learned systems, and the engineering of formal specification languages. Each supplies a specific tool or a specific caution.

On the dynamical side, the strongest precedent for substrate-independent control motifs is the recent treatment of habituation. [2] frames habituation not as a biological phenomenon but as a problem of designing a dynamical system: given qualitative constraints on transient response, identify the minimal dynamical ingredients that meet them, using a behavioral (Willems-style) formulation in which a system is defined by the set of admissible input-output trajectories $\mathcal{B}$ rather than by state-space equations. Two of its results carry directly into our setting. First, habituation has been observed not only in animals but in unicellular organisms and in non-living devices such as electronic circuits and neuromorphic materials, which is empirical evidence that a learned dynamical motif can recur across substrates with no shared biology. Second, the review proves that nonnegative, bounded habituation rules out linearity, making nonlinearity structural rather than optional, and that a single fading-memory state with a static nonlinear readout already captures the core hallmarks. Our universality criterion (Section 4) is the steering-side analogue of their synthesis question: which control operators have costs determined by structural invariants rather than by substrate detail.

Formal language theory has repeatedly been used to describe the behavior of learned neural systems, and each use exposes both the power and the limits of the approach. [3] surveys the connection from Chomsky’s hierarchy to modern NLP and records a caution that directly constrains our claims: constructions in which recurrent networks encode large computations exploit arbitrary numerical precision, such encodings are highly sensitive to slight perturbations of network parameters, and they correspond to steep valleys in parameter space that gradient-based optimization is unlikely to find. Any attractor calculus that assumes clean, well-separated basins in a trained network must therefore state how it behaves when basins are shallow or ill-conditioned; we address this in Section 6. [8] provides the closest methodological precedent on the descriptive side: the LEXR framework explains the decision-making of recurrent neural networks using Linear Temporal Logic, generates model-agnostic post-hoc local explanations inside a user-defined region of interest, carries a PAC-style guarantee on explanation accuracy, and

empirically produces explanations that are often smaller and easier to interpret than DFA-based extractions from RNN acceptors. LEXR shows that a formal language can be fitted to a trained recurrent system after the fact; our calculus asks the inverse question, namely which formal operators can be fitted before the fact as control commands. [11] quantifies what formal structure buys when used as an inductive bias during control: imposing formal-language constraints on Markov decision processes, combined with PPO and Lagrangian enforcement in Safety Gym, reduced the accumulated cost rate often by between almost one half and an order of magnitude, at the price of decreased performance in the earliest training steps while the constraint embedding was learned. That measured trade, formal structure purchased for a concrete cost reduction, is the pattern of benefit our calculus predicts for steering, and their observation that early training degrades before improving warns that operator-based steering may carry a transient cost of its own.

The engineering of formal specification languages supplies the second pillar: evidence that a small, well-chosen operator or component vocabulary can carry the weight of an entire specification practice. [5] describes the CBS framework, in which programming-language semantics are specified by translating programs to reusable constructs called funcons, whose semantics are defined once and for all in a modular variant of structural operational semantics; the framework exists precisely because scaling up full-language formal semantics by hand is usually regarded as a huge effort not worth undertaking. Our operator set is designed with the same reuse goal: each operator's cost theorem should be proved once and instantiated per substrate. [9] demonstrates the maturity such an approach can reach: Wasm SpecTec generates a formal specification, prose pseudocode, and mechanized definitions from a single source of truth written in a domain-specific language, and the extracted Wasm semantics passes all 23,778 applicable tests in the official Wasm test suite. SpecTec's stated success criterion, adoption by industrial stakeholders so that the normative definition is written in the DSL, is the adoption model we propose for an attractor-calculus standard in Section 7. On the mathematical side, [4] shows that a lightweight λ -calculus-style syntax (the μ -syntax) can represent an entire algebraic structure, the cyclic operad, compactly enough to give a complete step-by-step proof of the equivalence between its biased and unbiased definitions; this is a precedent for our claim that a small operator syntax can support nontrivial equivalence and cost proofs rather than merely serving as notation. [10] contributes an inheritance theorem of the shape we seek: a language-convexity criterion under which regularity of a language representation passes from a group to its finite-index subgroups, yielding Theorem 1.1, that a finitely generated group with a regular positive cone passes regularity to finite-index subgroups. The pattern, a criterion under which a property survives a change of substrate (group to subgroup), is exactly the logical form of our portability question (system to system). [7] illustrates how much structure a language class alone can carry: defining $L_n = \{w \in \Sigma^* : \xi(w) \text{ is divisible by } n\}$ with $\xi(w) = |w|_a - |w|_b$, the languages L_n are regular, and their union over primes equals $\{w : \xi(w) \neq \pm 1\}$, from which the infinitude of primes follows; we use the same discipline of deriving system-level facts from the closure properties of an operator language. [6] shows the graded expressiveness that formal automata families provide, with iterated pushdown automata sitting strictly between pushdown automata and Turing machines, and gives the Fibonacci language a^{f_n} as a concrete 2-iterated-pushdown example; this graded hierarchy is a reminder that our operator set should be classified by expressive power, not assumed maximal. Finally, [1] constructs reflectica, a formal language that is its own meta-language, in which every object, including expressions and their values, is a finite string with a canonical name; a calculus of steering that is intended to describe systems which themselves process the calculus' tokens (such as language models prompted in the calculus' own notation) will eventually need exactly this self-referential capacity, and we flag it as future work rather than claiming it here.

3. Methods

3.1 State-space semantics

We assume every information-processing system is modeled as a controlled dynamical system

$$\dot{x} = f(x, u) + \sigma\xi(t), \quad x \in \mathbb{R}^n,$$

where x is the internal state, $u \in \mathcal{U}$ is the control signal (a prompt, a stimulus, a stimulation current), $\sigma\xi(t)$ is additive noise of intensity $D = \sigma^2/2$, and the uncontrolled flow $f(x, 0)$ possesses a set of attracting equilibria or attracting sets $\{A_i\}$, the attractors. The basin B_i of A_i is the set of initial conditions whose trajectories converge to A_i . For the analytical results of this paper we work in one dimension with a bistable gradient system

$$\dot{x} = -V'(x) + u, \quad V(x) = \frac{x^4}{4} - \frac{x^2}{2},$$

which has stable equilibria at $x = \pm 1$ with $V(\pm 1) = -1/4$ and a saddle at $x = 0$ with $V(0) = 0$. This potential is the canonical minimal model of two competing attractor basins separated by a barrier; every derivation in Section 4 states its inputs explicitly, and all of them are model choices or labeled assumptions, not measurements.

3.2 Cost functional

The cost of a control signal $u : [0, T] \rightarrow \mathcal{U}$ is the quadratic energy

$$E[u] = \int_0^T \|u(t)\|^2 dt.$$

For an impulse $u(t) = p\delta(t)$, this gives $E[u] = p^2$. The quadratic functional is chosen because it is the standard minimal-energy criterion in control theory and because it makes the composition algebra in Section 3.4 additive; we examine the sensitivity of the results to this choice in Section 6.

3.3 Operator set

The calculus consists of five operators, each mapping a system state and a goal to a control signal:

- **PERTURB**($A_i \rightarrow A_j$): deliver a minimal impulse that moves the state from basin B_i across the separatrix into B_j .
- **HOLD**(A_j, T_d): maintain the state inside B_j for a dwell time T_d against noise, typically by a sustaining input or by prior basin deepening.
- **DEEPEN_BASIN**($A_j, \delta d$): modify the system (or apply a sustained background input) so that the effective barrier protecting A_j increases by δd .
- **SHIFT_BOUNDARY**($A_i \rightarrow A_j$): apply a sustained input that tilts the landscape until the separatrix moves and the state falls into B_j without an impulse.
- **CREATE_ATTRACTOR**(A_{new}): introduce a parameter change that produces a new attracting equilibrium where none existed.

For this paper, **CREATE_ATTRACTOR** is defined but not costed: its minimal cost depends on global bifurcation structure that the one-dimensional model does not exhibit, and we state it as an open problem.

3.4 Composition algebra

Operators compose sequentially. If u_1 is supported on $[0, T_1]$ and u_2 on $[T_1, T_1 + T_2]$, then

$$E[u_1 \cdot u_2] = \int_0^{T_1} \|u_1\|^2 dt + \int_{T_1}^{T_1+T_2} \|u_2\|^2 dt = E[u_1] + E[u_2],$$

so the cost functional is additive over sequential composition with disjoint time supports. This additivity is the algebraic backbone of the calculus: it lets a compiled steering program be costed by summing its operators' costs, exactly as a funcon-term program in [5] is given semantics by the once-and-for-all semantics of its components.

3.5 Universality criterion

We define an operator's cost function c_{op} to be portable across a class of substrates if c_{op} is a function only of invariant quantities of the attractor landscape, such as the barrier height ΔV , the noise intensity D , and basin volumes, and not of substrate-specific quantities such as the local curvature profile of the barrier or the parameterization of the control channel. This criterion is deliberately analogous to the inheritance criterion of [10]: language-convexity there, cost-invariance here, both asking when a property survives a change of carrier.

4. Analysis

Every input number in this section is either a property of the model potential defined in Section 3.1 (a stated model choice) or an explicitly labeled assumption. All arithmetic is shown.

4.1 Barrier height of the canonical potential

With $V(x) = x^4/4 - x^2/2$: the stationary points satisfy $V'(x) = x^3 - x = 0$, giving $x \in \{-1, 0, 1\}$. The minima are at $x = \pm 1$:

$$V(\pm 1) = \frac{1}{4} - \frac{1}{2} = -\frac{1}{4},$$

and the saddle is at $x = 0$ with $V(0) = 0$. The barrier height separating the two basins is therefore

$$\Delta V = V(0) - V(\pm 1) = 0 - \left(-\frac{1}{4}\right) = \frac{1}{4} = 0.25.$$

This value $\Delta V = 0.25$ is the single invariant on which the impulse results below rest.

4.2 PERTURB: minimal impulse cost

We adopt the following convention for the impulse channel, stated here as a model choice: the control input acts as a force on an effectively underdamped steering coordinate, so that an impulse $u(t) = p\delta(t)$ delivers momentum p to the state and the quadratic cost functional of Section 3.2 gives $E[u] = p^2$. Crossing from basin B_- into B_+ requires the post-impulse energy to exceed the barrier: the kinetic energy $p^2/2$ delivered at the minimum $x = -1$ must satisfy

$$\frac{p^2}{2} \geq \Delta V.$$

The minimal impulse therefore has $p^2 = 2\Delta V$, and the minimal cost is

$$E_{\text{imp}} = 2\Delta V = 2 \cdot \frac{1}{4} = \frac{1}{2} = 0.5.$$

The corresponding impulse amplitude is $p = \sqrt{2\Delta V} = \sqrt{0.5} \approx 0.707107$. The qualitative content of this result is not the constant 0.5 but its dependence structure: the cost is a function of the invariant ΔV alone, and no other property of the potential enters.

4.3 SHIFT_BOUNDARY: minimal constant-tilt cost

Now let the control be a constant input $u(t) = u$ sustained over $[0, T]$. The controlled system is $\dot{x} = x - x^3 + u$; equilibria satisfy $x^3 - x = u$. Starting in the left well, the state remains trapped until the left minimum and the saddle annihilate in a saddle-node bifurcation, after which the state rolls into B_+ . The annihilation occurs when the cubic $x^3 - x - u = 0$ has a double root, i.e., when both the equilibrium condition and its derivative vanish:

$$x^3 - x = u, \quad 3x^2 - 1 = 0.$$

The relevant double root on the left branch is $x = -1/\sqrt{3}$, giving the critical tilt

$$u_c = \left(-\frac{1}{\sqrt{3}}\right)^3 - \left(-\frac{1}{\sqrt{3}}\right) = -\frac{1}{3\sqrt{3}} + \frac{1}{\sqrt{3}} = \frac{2}{3\sqrt{3}} \approx 0.384900.$$

The cost rate of the minimal sustained input is

$$u_c^2 = \frac{4}{27} \approx 0.148148 \text{ per unit time,}$$

and the total cost of a tilt held for time T is $E_{\text{tilt}}(T) = \frac{4}{27}T$.

4.4 Break-even horizon

The impulse strategy costs $E_{\text{imp}} = 0.5$ once; the tilt strategy costs $\frac{4}{27}T$. The two strategies break even at the horizon T^* satisfying

$$\frac{4}{27}T^* = 0.5 \quad \implies \quad T^* = \frac{0.5 \cdot 27}{4} = \frac{27}{8} = 3.375.$$

For steering tasks shorter than $T^* = 3.375$ time units the tilt is cheaper; for longer tasks the impulse is cheaper. This is the first compiled decision rule of the calculus: given a dwell-time requirement, the choice between PERTURB and SHIFT_BOUNDARY is decided by comparing the required dwell time against T^* .

4.5 Generalization to $V(x) = x^4/4 - ax^2/2$

For the one-parameter family $V_a(x) = x^4/4 - ax^2/2$ with $a > 0$: the minima sit at $x = \pm\sqrt{a}$ with

$$V_a(\pm\sqrt{a}) = \frac{a^2}{4} - \frac{a^2}{2} = -\frac{a^2}{4},$$

the saddle at $x = 0$ has $V_a(0) = 0$, and the barrier height is

$$\Delta V(a) = \frac{a^2}{4}.$$

The impulse cost generalizes to

$$E_{\text{imp}}(a) = 2\Delta V(a) = \frac{a^2}{2}.$$

For the tilt, the equilibrium condition is $x^3 - ax = u$ and the double-root condition is $3x^2 - a = 0$, giving $x = -\sqrt{a/3}$ and

$$u_c(a) = \left(-\sqrt{\frac{a}{3}}\right)^3 - a\left(-\sqrt{\frac{a}{3}}\right) = -\frac{a}{3}\sqrt{\frac{a}{3}} + a\sqrt{\frac{a}{3}} = \frac{2a}{3}\sqrt{\frac{a}{3}} = \frac{2a^{3/2}}{3\sqrt{3}},$$

so the tilt cost rate is

$$u_c^2(a) = \frac{4a^3}{27}.$$

The break-even horizon follows by equating $\frac{4a^3}{27}T$ with $\frac{a^2}{2}$:

$$T^*(a) = \frac{a^2/2}{4a^3/27} = \frac{27}{8a}.$$

4.6 HOLD: noise-robustness and required basin depth

The HOLD operator must keep the state inside a basin for a dwell time T_d against noise of intensity D . We adopt the standard Kramers-type escape model as a labeled assumption: the mean escape rate from a basin protected by a barrier of height ΔV under noise intensity D is

$$r \approx r_0 e^{-\Delta V/D},$$

and we set the attempt prefactor $r_0 = 1$ per unit time, which is the most conservative choice in the sense that it maximizes the predicted escape rate for any given ΔV ; every number below inherits this assumption. The probability of at least one escape during the dwell time is then, to first order in rT_d ,

$$P_{\text{fail}} \approx 1 - e^{-rT_d} \approx rT_d.$$

Requiring $P_{\text{fail}} \leq \varepsilon$ gives the design rule

$$rT_d \leq -\ln(1 - \varepsilon) \implies \Delta V' \geq D \ln\left(\frac{T_d}{-\ln(1 - \varepsilon)}\right).$$

We now evaluate this at the operating point stated in the Abstract: $D = 0.05$, $\varepsilon = 10^{-3}$, $T_d = 10^6$ time units. The arithmetic is:

$$-\ln(1 - 10^{-3}) = -\ln(0.999) \approx 0.0010005,$$

$$\frac{T_d}{-\ln(1 - \varepsilon)} = \frac{10^6}{0.0010005} \approx 9.99500 \times 10^8,$$

$$\ln(9.99500 \times 10^8) = \ln(10^9) + \ln(0.99950) \approx 20.72327 - 0.00050 \approx 20.722766,$$

$$\Delta V' \geq 0.05 \times 20.722766 \approx 1.0361383.$$

The canonical potential already provides $\Delta V = 0.25$, so the additional depth that DEEPEN_BASIN must supply is

$$\delta d = \Delta V' - \Delta V \approx 1.0361383 - 0.25 = 0.7861383 \approx 0.786138.$$

Two cost consequences follow immediately from the additivity of Section 3.4 and the impulse theorem of Section 4.2. First, if the deepening is done before an impulse crossing out of the deepened basin, the impulse cost rises to

$$E'_{\text{imp}} = 2\Delta V' \approx 2 \times 1.0361383 = 2.072277,$$

an increase of $2\delta d \approx 1.572277$ over the undeepened cost of 0.5. Second, the break-even horizon between impulse and tilt shifts, because the tilt cost rate $u_c^2 = 4/27$ is unchanged by deepening (the tilt operator removes the barrier rather than climbing it) while the impulse cost grows:

$$T^{*'} = \frac{E'_{\text{imp}}}{u_c^2} = \frac{2.072277}{4/27} = 2.072277 \times \frac{27}{4} \approx 13.987867.$$

So a reliability requirement roughly quadruples the horizon ($3.375 \rightarrow 13.987867$) within which the tilt strategy is the cheaper crossing primitive.

4.7 Universality classification of the operator set

Applying the criterion of Section 3.5 to the results derived above yields a partial classification:

- **PERTURB is portable.** Its minimal cost $E_{\text{imp}} = 2\Delta V$ is a function of the invariant barrier height alone (Section 4.2), and the generalization $E_{\text{imp}}(a) = a^2/2$ (Section 4.5) confirms that no curvature or parameterization detail enters. Two substrates with equal ΔV have equal PERTURB cost by construction.
- **SHIFT_BOUNDARY is substrate-specific.** Its cost rate $u_c^2(a) = 4a^3/27$ depends on the parameter a through the local geometry of the barrier top, not only through $\Delta V(a) = a^2/4$: two potentials with the same barrier height but different a (and hence different well separation and curvature) have different tilt costs. Concretely, $a = 1$ gives $u_c^2 = 4/27 \approx 0.148148$, while $a = 2$ gives $\Delta V(2) = 1$ — the same as a potential with a chosen so that $\Delta V = 1$ in another family — but $u_c^2(2) = 32/27 \approx 1.185185$, an 8-fold increase in cost rate for a 4-fold increase in barrier height. The ratio $u_c^2/\Delta V = (4a^3/27)/(a^2/4) = 16a/27$ is not invariant; it scales linearly with a .
- **DEEPEN_BASIN and HOLD are conditionally portable.** The required depth $\Delta V' \geq D \ln(T_d/(-\ln(1 - \varepsilon)))$ is a function of the invariants D, T_d, ε only, so the *target* is portable; but the *cost* of achieving a depth increase δd depends entirely on the actuation mechanism (parameter change, sustained input, weight change in a network), which is substrate-specific and not costed here.
- **CREATE_ATTRACTOR is unclassified.** Its cost depends on global bifurcation structure outside the one-dimensional model; it is an open problem (Section 3.3).

The classification is partial but sharp: the calculus predicts that impulse-like steering primitives should transfer across substrates with cost rescaled by ΔV , while landscape-reshaping primitives must be re-derived per system. This is the testable content of the universality conjecture at the level of costs.

5. Results

All numbers below are derived in Section 4 with shown arithmetic; none is measured. They are properties of the canonical model and the labeled assumptions of Sections 4.2 and 4.6.

Quantity	Symbol	Value	Source
Barrier height	ΔV	0.25	Section 4.1
Minimal impulse cost	E_{imp}	$2\Delta V = 0.5$	Section 4.2
Impulse amplitude	p	$\sqrt{0.5} \approx 0.707107$	Section 4.2
Critical tilt	u_c	$2/(3\sqrt{3}) \approx 0.384900$	Section 4.3
Tilt cost rate	u_c^2	$4/27 \approx 0.148148$	Section 4.3

Quantity	Symbol	Value	Source
Break-even horizon	T^*	$27/8 = 3.375$	Section 4.4
Generalized barrier height	$\Delta V(a)$	$a^2/4$	Section 4.5
Generalized impulse cost	$E_{\text{imp}}(a)$	$a^2/2$	Section 4.5
Generalized tilt cost rate	$u_c^2(a)$	$4a^3/27$	Section 4.5
Generalized break-even horizon	$T^*(a)$	$27/(8a)$	Section 4.5
Required deepened barrier	$\Delta V'$	≥ 1.0361383	Section 4.6
Required depth increase	δd	≈ 0.786138	Section 4.6
Deepened impulse cost	E'_{imp}	≈ 2.072277	Section 4.6
Deepened break-even horizon	$T^{*'} $	≈ 13.987867	Section 4.6

The principal results are three. First, the minimal-energy costs of the two crossing primitives admit closed forms, $E_{\text{imp}} = 2\Delta V$ and $E_{\text{tilt}}(T) = (4/27)T$, with a break-even horizon $T^* = 3.375$ that compiles the PERTURB-vs-SHIFTBOUNDARY choice into a single comparison against the required dwell time. Second, a reliability requirement of failure probability $\leq 10^{-3}$ over $T_d = 10^6$ time units at $D = 0.05$ forces $\delta d \approx 0.786138$ of additional barrier depth, raising impulse cost by $2\delta d \approx 1.572277$ and the break-even horizon to ≈ 13.987870 . Third, the operator set splits under the portability criterion: PERTURB cost is invariant-determined and portable; SHIFTBOUNDARY cost is geometry-determined and substrate-specific; HOLD/DEEPEN_BASIN have portable targets but substrate-specific costs.

6. Discussion

The conjecture is not tested here. Every result in this paper is a theorem about a one-dimensional model potential plus labeled assumptions; no experiment on a trained network, an LLM, or a neural circuit is reported. The correct reading is that the calculus makes a specific, falsifiable prediction about *cost structure*, and the empirical program of the research idea (minimal-perturbation steering in toy recurrent networks and trained LLMs, with transfer tests between architectures) is the test. What the formal layer contributes is the prediction to check: if cross-substrate transfer of steering primitives exists, PERTURB-like operators should transfer with costs rescaling by ΔV , and SHIFT_BOUNDARY-like operators should not.

Limitations. The results rest on the following assumptions, each with a stated sensitivity:

- *Canonical potential.* All closed forms are for $V(x) = x^4/4 - x^2/2$ and its one-parameter family. The constants 0.5, $4/27$, 3.375 change with the potential; only the dependence structure (impulse cost $\propto \Delta V$; tilt cost $\propto$ barrier-top geometry) is claimed to generalize. Sensitivity: high for the numerical values, low for the classification.
- *Impulse convention.* Section 4.2 assumes an underdamped steering coordinate in which an impulse delivers momentum p with cost p^2 . In a strongly overdamped coordinate, crossing requires sustained force over a finite time and the impulse cost formula fails. Sensitivity: high; this is the single most load-bearing modeling choice.
- *Kramers prefactor.* Section 4.6 sets $r_0 = 1$ per unit time. If the true prefactor is $r_0 = 10^2$, the required depth increases by $D \ln(100) = 0.05 \times 4.60517 \approx 0.230259$; if $r_0 = 10^{-2}$, it decreases by the same amount. Sensitivity: moderate, logarithmic in r_0 .
- *Quadratic cost functional.* The additivity of Section 3.4 and the impulse cost both depend on the quadratic choice. An L^1 cost would change the minimal-impulse solution and break the clean additivity. Sensitivity:

moderate for costs, low for the portability classification, which concerns dependence structure rather than cost values.

- *One dimension, additive noise.* Real high-dimensional systems have saddle escape along specific unstable directions, and noise may be multiplicative and state-dependent. The one-dimensional analysis is the $n = 1$ case of the semantics, not a reduction of the general case.
- *No empirical content yet.* The claim that LLMs or brains are attractor-based dynamical systems in the sense of Section 3.1 is a conjecture, not a finding.

What would falsify the main result. The main result is the cost-structure split: PERTURB cost a function of ΔV alone, SHIFTBOUNDARY cost geometry-dependent. *The single observation that would falsify it is a pair of systems (or two parameter settings of one system) with equal measured barrier height ΔV but systematically different minimal impulse costs for the same crossing reliability — that is, a measured impulse cost that depends on something other than ΔV . Conversely, if tilt-style sustained-input costs were found to be predictable from ΔV alone across substrates, the claimed substrate-specificity of SHIFTBOUNDARY would be falsified.* A third falsifier targets the framework itself: if trained networks subjected to minimal-perturbation steering show no metastable regime structure at all (no bistability, no hysteresis, no dwell-time statistics consistent with barrier crossing), the attractor semantics of Section 3.1 does not apply and the calculus is vacuous for those systems.

Failure modes. Three are foreseeable. First, shallow or ill-conditioned basins: as [3] records, recurrent-network constructions that encode large computations exploit arbitrary numerical precision, are highly sensitive to slight parameter perturbations, and correspond to steep valleys that gradient-based optimization is unlikely to find; a calculus that assumes clean basins will mispredict costs exactly where trained networks are most likely to live. Second, transient steering costs: [11] measured that imposing formal-language constraints during reinforcement learning degraded performance in the earliest training steps before improving; operator-based steering may analogously carry a transient destabilization not captured by the steady-state cost functional.

Open questions. (i) A cost theorem for CREATE_ATTRACTOR, which requires at least a two-parameter model exhibiting a saddle-node or pitchfork creation bifurcation. (ii) The self-referential extension: a calculus whose tokens can themselves be prompts to the system it steers will need the self-reflection capacity that [1] builds into reflectica, where every object, including expressions and their values, is a finite string with a canonical name; this is flagged, not solved. (iii) Whether the graded expressiveness hierarchy of automata families described in [6] — where iterated pushdown automata sit strictly between pushdown automata and Turing machines — offers the right complexity grading for operator languages over high-dimensional basins.

7. Practical Implications and Engineering Pathways

(a) Who would use this and what changes. The immediate users, if the conjecture survives testing, are three groups. Prompt engineers working on large language models would gain a cost model: instead of iterating on wording until outputs improve, they would state a steering goal as an operator sequence (e.g., DEEPEN_BASIN on an instruction-following regime, then PERTURB across a topic boundary) and predict the intervention budget before running it. Neuromodulation and neurofeedback protocol designers would gain a shared vocabulary with the prompting community, making transfer claims between stimulation protocols and computational models stateable and checkable rather than analogical. Behavioral intervention designers would gain an audit tool: a stimulus sequence is doing semantic work only if its cost profile matches the predicted operator costs, not merely if it shifts outputs. What changes concretely is the unit of reuse: control primitives become portable artifacts with declared cost invariants, rather than per-system tricks. At present no user can apply the calculus, because no cost theorem has been validated on a real system; the work is foundational in the sense that it supplies the language and the predictions that the first validation experiments would test.

(b) Readiness level. This work is purely analytical and has not been peer reviewed: by the NASA Technology Readiness definitions it sits at TRL 1–2 (basic principles observed and reported; concept and application formulated). Analytical work does not exceed TRL 3, and nothing here approaches experimental proof of concept on a real system. No manufacturing or deployment readiness is claimed.

(c) The smallest next experiment. The cheapest de-risking test is on a trained recurrent network, not an LLM. Protocol: train a small recurrent network on a task with two known metastable output regimes; estimate the effective barrier height ΔV between regimes from dwell-time statistics under noise (the escape-rate fit of Section 4.6 gives ΔV from $-\ln r$ vs. $1/D$); then deliver minimal-norm perturbations to cross between regimes and measure the perturbation norm at a fixed success rate. The prediction to test is that the measured minimal perturbation energy scales linearly with the independently estimated ΔV across a family of networks or noise levels. The outcome that falsifies the portable-operator claim is a measured impulse cost that does not track ΔV — for example, two networks with equal dwell-time-inferred ΔV but a several-fold difference in minimal crossing energy. A secondary measurement, the tilt analogue, applies a sustained bias input and measures the critical bias at which the regime flips; the prediction is that this critical bias does *not* collapse across networks with equal ΔV , confirming the substrate-specific classification.

(d) Binding constraints. In order of likely first contact: (i) *measurement of ΔV itself* — dwell-time statistics require many long trajectories per network, and the Kramers fit assumes a prefactor that must itself be estimated, so the invariant the whole cost theory rests on is the hardest quantity to measure; (ii) *basin quality in trained networks* — the sensitivity of precision-based constructions to parameter perturbation documented in [3] means real basins may be shallow and the closed forms may hold only in a narrow parameter window; (iii) *control-channel parameterization* — the impulse convention of Section 4.2 presumes an underdamped steering coordinate, and identifying which internal directions of a trained network admit such a coordinate is an unsolved system-identification problem; (iv) *compute for the LLM case* — perturbation sweeps over a large model's activations are cheap per query but require many queries, and the transfer test across architectures (transformer vs. RNN) doubles the sweep. Cost figures for these constraints are not derivable from the present analysis and are left as open questions.

(e) Order-of-magnitude projections. One projection is derivable from Section 4 and is labeled as such. If a steering task requires dwell time T_d at failure probability ε under noise D , the deepened impulse cost is $E'_{\text{imp}} = 2D \ln(T_d/(-\ln(1 - \varepsilon)))$, which grows only logarithmically in T_d : at the assumed operating point, increasing the dwell requirement from 10^6 to 10^9 time units adds $2D \ln(10^3) = 2 \times 0.05 \times 6.90776 \approx 0.690776$ to the impulse cost. The practical reading is that reliability requirements are cheap in energy but expensive in required basin depth, which is the quantity substrate owners (model trainers, device manufacturers) control. All other cost figures — per-query budgets for LLM steering, stimulation amplitudes for neural protocols — depend on substrate parameters not derived here and are not projected.

(f) Risks and misuse. A validated cost calculus for steering latent states is dual-use. The same operator language that compiles a therapeutic stimulation protocol compiles an undetectable steering sequence: we adopt as a working hypothesis, not a finding of this paper, that non-semantic prompt features can shift model latent states; a cost calculus would make such shifts cheaper and more systematic. An audit capability (does this stimulus sequence do semantic work?) is therefore not optional but a required companion of the steering capability, and any standardization effort along the lines of the adoption model in Section 7(a) should carry the audit operators in the same specification.

8. Conclusion

This paper built the formal layer that makes the substrate-independence conjecture about cognitive steering testable, and derived its first closed-form consequences. The Attractor Calculus defines five operators over attractor-based dynamical systems with a quadratic, additive cost functional; for the canonical bistable potential it yields minimal impulse cost $E_{\text{imp}} = 2\Delta V = 0.5$, tilt cost rate $u_c^2 = 4/27 \approx 0.148148$, break-even horizon $T^* = 27/8 = 3.375$, and, under the stated reliability operating point, a required basin deepening of $\delta d \approx 0.786138$ that raises impulse cost by ≈ 1.572277 and the break-even horizon to ≈ 13.987870 . The principal structural result is the partial universality classification: impulse-operator cost depends only on the invariant barrier height and is therefore portable in the defined sense, while tilt-operator cost depends on barrier-top geometry and is substrate-specific. This split is the calculus' first falsifiable prediction, and the smallest experiment that tests it — dwell-time-inferred ΔV versus measured minimal crossing energy across trained recurrent networks — is specified in Section 7. If the prediction holds, steering primitives become portable artifacts with declared cost invariants across prompting, stimulation, and intervention; if it fails, cross-substrate transfer claims in those fields lose their warrant. Either outcome is informative, and the formal layer built here is what makes either outcome readable.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.18.6-codeagent). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, zai-org/glm-5.3, openai/gpt-oss-120b, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: a separately generated script recomputed 16 quantitative claims from the paper's stated inputs; 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

- [1] A Self-Reflecting Formal Language. arXiv:2302.09077. <https://arxiv.org/abs/2302.09077> [2] Dynamical principles of habituation across substrates and scales. arXiv:2608.00249. <https://arxiv.org/abs/2608.00249> [3] Formal Language Theory Meets Modern NLP. arXiv:2102.10094. <https://arxiv.org/abs/2102.10094> [4] A formal language for cyclic operads. arXiv:1602.07502. <https://arxiv.org/abs/1602.07502> [5] A Component-Based Formal Language Workbench. arXiv:1912.10631. <https://arxiv.org/abs/1912.10631> [6] Hyperbolic tilings and formal language theory. arXiv:1309.1271. <https://arxiv.org/abs/1309.1271> [7] Infinitude of Primes Using Formal Language

Theory. arXiv:2005.10372. <https://arxiv.org/abs/2005.10372> [8] A Formal Language Approach to Explaining RNNs. arXiv:2006.07292. <https://arxiv.org/abs/2006.07292> [9] Wasm SpecTec: Engineering a Formal Language Standard. arXiv:2311.07223. <https://arxiv.org/abs/2311.07223> [10] Formal language convexity in left-orderable groups. arXiv:1905.13001. <https://arxiv.org/abs/1905.13001> [11] Formal Language Constraints for Markov Decision Processes. arXiv:1910.01074. <https://arxiv.org/abs/1910.01074>