Attention as Active Stochastic Control: A Unifying Theory and Architectural Framework

Anonymous authors
Paper under double-blind review

Abstract

For nearly a decade, the Transformer architecture has served as the empirical workhorse of modern artificial intelligence. Despite its ubiquitous deployment, its core mechanism, softmax self-attention, is primarily understood as a heuristic pairwise sequence mixer optimized via gradient descent. Consequently, pervasive failure modes such as prompt fragility, factual hallucinations, representation collapse (over-smoothing), and attention sinks have been addressed largely through ad-hoc engineering heuristics. In this work, we establish a rigorous mathematical unification between Transformer Attention and Active Stochastic Control. Specifically, we demonstrate that along the depth axis, self-attention operates as an active observer executing dual control: the query functions as an exploratory control action steering sensor admittance to minimize epistemic entropy, the key matrix defines a directional channel alignment metric, and the softmax operator arises as the exact analytical Bayesian update of a Wonham filter for categorical point-process emissions. Furthermore, viewing residual stacks as continuous-time Hamiltonian interacting particle systems exposes a symplectic leapfrog structure enabling $\mathcal{O}(1)$ activation memory training. Leveraging this framework, we introduce Closed-Loop Iterated Active Attention (IAA), which resolves multi-hop relational reasoning within a single physical layer, reuses a static key-value observation manifold, and terminates computation dynamically via Kalman innovation norm decay. Finally, we translate foundational principles from robust control (H_∞), balanced truncation (Hankel singular values), and Bode's sensitivity integral to provide principled explanations and actionable solutions for transformer jailbreaks, KV-cache compression, and attention sinks.

1 Introduction

Modern sequence modeling is dominated by the Transformer architecture (Vaswani et al., 2017). While scaling laws have demonstrated consistent empirical returns across model width, depth, and dataset cardinality, foundational architectural questions remain open:

1. Why is softmax self-attention so remarkably effective? Attention provides a powerful form of content-addressable retrieval, yet it remains an open question whether its particular softmax formulation is merely an effective heuristic mixer or reflects a foundational principle of optimal stochastic inference.
2. Why is iterative multi-step reasoning computationally inefficient? Many relational and algorithmic tasks require computation whose effective depth grows with problem complexity. Standard fixed-depth Transformers cannot dynamically adjust internal sequential computation, forcing iterative reasoning into autoregressive Chain-of-Thought (CoT) generation and redundant token-by-token forward passes.
3. What explains persistent architectural pathologies? Standard models exhibit well-documented empirical anomalies, including attention sinks (Xiao et al., 2023) (where early tokens absorb disproportionate attention mass), susceptibility to adversarial suffixes, and performance degradation over long, compositional reasoning chains. A unified, first-principles explanation for these behaviors remains incomplete.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In this paper, we resolve these questions by establishing that the Transformer is an Active Stochastic Observer operating across depth. While prior literature has explored connections between self-attention and continuous Hopfield networks or dense associative memories (Hopfield, 1982; Krotov and Hopfield, 2016; Demircigil et al., 2017; Ramsauer et al., 2020), as well as external neural memory architectures (Graves et al., 2014; Weston et al., 2014; Ba et al., 2016), these formulations largely treat attention as a passive static energy retrieval mechanism. By bridging modern attention mechanisms with classical estimation theory (Kalman, 1960; Wonham, 1965), dual control (Feldbaum, 1965; Friston, 2010), and robust state-space control (Bode, 1945; Doyle et al., 1989; Zhou et al., 1996), we construct an analytical bridge that transforms empirical deep learning tricks into predictable control-theoretic engineering laws.

The remainder of this paper is organized as follows: Section 2 establishes our foundational theory, unifying asymmetric directional admittance, dual control teleology, and the continuous-time Wonham filter to prove that standard attention is an open-loop stochastic observer. Section 3 introduces Closed-Loop Iterated Active Attention (IAA), closing the estimation loop via internal query saccades to resolve multi-hop relational dependencies within a single physical layer. Section 4 scales this framework to macro-scale network depth, modeling the residual backbone as a continuous Hamiltonian dynamical system that achieves strictly $\mathcal{O}(1)$ activation memory via symplectic leapfrog integration. Section 5 develops a system-theoretic diagnostic suite, translating classical control invariants into predictable laws for attention sinks, KV-cache compression, and adversarial stability. Finally, Section 6 provides comprehensive empirical validation across synthetic multi-hop reasoning tasks and open foundation models.

2 Self-Attention as an Active Stochastic Observer

2.1 Directional Channel Admittance and Active Dual Control

To understand why softmax attention serves as such an effective inferential mechanism, we begin by examining its unnormalized scoring geometry at the individual token level. Standard literature frequently conceptualizes attention as computing pairwise similarity or content-based addressing across an embedding metric space. However, inspecting the bilinear algebraic structure of the query-key interaction reveals that self-attention does not implement a symmetric distance metric, but rather a directed sensing operator.

In standard self-attention, input tokens $X \in \mathbb{R}^{N\times d}$ are projected to queries $Q = XW_Q$ and keys $K = XW_K$, where $W_Q, W_K \in \mathbb{R}^{d\times d_k}$. The unnormalized pairwise score matrix expands as an asymmetric bilinear form:

$$S = \frac{1}{\sqrt{d_k}} QK^T = \frac{1}{\sqrt{d_k}} X(W_Q W_K^T)X^T \tag{1}$$

While standard token Gram matrices XX^T are inherently symmetric and positive semi-definite, the learned bilinear operator $M = W_Q W_K^T \in \mathbb{R}^{d\times d}$ is non-symmetric whenever $W_Q \neq W_K$. The scalar entry:

$$S_{ij} = \frac{1}{\sqrt{d_k}} q_i^T k_j = \frac{1}{\sqrt{d_k}} x_i^T M x_j \neq S_{ji} \tag{2}$$

does not represent a static auto-covariance between random vectors, but rather the asymmetric channel admittance between the interrogation probe issued by observer token i and the semantic measurement channel offered by token j. This structure captures an essential property of active sensing: information harvesting is non-reciprocal, as the evidential value of token j for updating token i's belief depends entirely on the observer's current epistemic state.

This adaptive sensing aperture bridges naturally to classical Dual Control (Feldbaum, 1965). In passive stochastic estimation, measurements $z_t = Hx_t + v_t$ are observed through fixed physical channels. In Active Sensing and Dual Control, the controller issues control actions u_t that serve two distinct objectives: (1) Control / Enacting Goal: driving the plant state

Control Theory Component	Transformer Sub-System	Functional Teleology
Belief State $\hat{x}_t^-$	Residual Stream $x_i^{(l)}$	Sufficient statistic of context history.
Measurement Field $\mathcal{Z}$	Sequence Tokens $X = [x_1, \ldots, x_N]$	Contextual environment for interrogation.
Probing Control Action u_t^s	Query Vector $q_i = x_i W_Q$	Action steering sensor aperture.
Sensor Admittance $H(u)$	Key Matrix $K = XW_K$	Directional channel alignment metric.
Measurement Noise Scale R	Temperature $\sqrt{d_k}$	Normalizer preventing certainty saturation.
Filter Gain K_t	Attention Matrix $A_{i,:} = \text{softmax}(S_{i,:})$	Data-dependent gain weighting measurements.
Measurement Payload z_t	Value Projection $V = XW_V$	Extrinsic feature signal harvested.
Innovation Update $\Delta\hat{x}_t$	Attention Output $\Delta x_i = A_{i,:}VW_O$	Assimilation of external evidence.
Transition Dynamics $f(\hat{x})$	Feed-Forward Network (FFN)	Associative key-value memory retrieval.
Posterior Belief State $\hat{x}_t^+$	Residual Update $x_i^{(l+1)}$	Updated belief state for layer $l+1$.

Table 1: Exact structural mapping between Active Stochastic Control and Transformer components.

toward an operational objective; and (2) Identification / Probing Goal: modulating the sensor admittance matrix $H(u_t)$ to reduce epistemic uncertainty over unobserved states. Within each Transformer layer l, the residual stream vector $x_i^{(l)}$ constitutes the current belief state of token i, while the query vector $q_i = x_i W_Q$ acts as an active probing action steering the sensor aperture against contextual keys. Table 1 details the exact structural mapping linking control theory components to Transformer sub-systems.

2.2 The Discrete Lexicon and the Wonham Categorical Filter

Having established that the residual stream maintains a latent belief state and the query acts as a directional probing control, we must determine the mathematically optimal Bayesian filter for assimilating contextual evidence. In standard estimation theory, continuous observations subject to Gaussian noise yield the linear Kalman filter. Natural language, however, does not arrive as continuous Gaussian diffusions; it consists of discrete categorical emissions sampled from a finite vocabulary $\mathcal{V}$ ($|\mathcal{V}| \approx 32\text{k} - 128\text{k}$).

Proposition 1 (Wonham Categorical Filter Equivalence). Let a continuous semantic state vector $x \in \mathbb{R}^d$ drive a discrete Poisson jump process with rates conditioned on categorical events $j \in \{1, \ldots, N\}$. The optimal minimum variance filter updating the posterior probability distribution $\pi_j = P(\text{token } j \mid x)$ is analytically governed by the Wonham filter (Wonham, 1965). Its discrete-time likelihood update is given precisely by the softmax function:

$$\pi_j = \frac{\exp(q^T k_j / \sqrt{d_k})}{\sum_{m=1}^{N} \exp(q^T k_m / \sqrt{d_k})} \tag{3}$$

While Softmax is conventionally introduced as an empirical differentiable argmax or a heuristic log-linear link function, Proposition 1 establishes that it emerges from optimal minimum-variance state estimation over discrete lexicons. Specifically, when continuous latent semantics drive marked point-process arrivals, the continuous nonlinear filtering problem is governed by the Kushner-Stratonovich stochastic differential equation. In Appendix A, we provide the complete continuous-to-discrete derivation, demonstrating that freezing latent transitions over an observation epoch yields the Softmax self-attention update as its unique analytical discrete likelihood assimilation, with $\sqrt{d_k}$ emerging naturally as the sensor measurement noise variance scale.

Closed-Loop Iterated Active Attention Architecture

Input: Static cached Keys $K = XW_K \in \mathbb{R}^{N \times d_k}$ and Values $V = XW_V \in \mathbb{R}^{N \times d_v}$
Initialize Query: $q_i^{(0)} = x_i W_Q \in \mathbb{R}^{d_k}$
For $\tau = 0, \ldots, \tau_{\max} - 1$:
$A_{i,:}^{(\tau)} = \text{softmax}\left(q_i^{(\tau)} K^T / \sqrt{d_k}\right) \in \Delta^{N-1}$
$v_{\text{agg}}^{(\tau)} = A_{i,:}^{(\tau)} V \in \mathbb{R}^{d_v}$
$\Delta q_i^{(\tau)} = W_{\text{update}} v_{\text{agg}}^{(\tau)}$ (Epistemic Kalman Innovation)
if $\|\Delta q_i^{(\tau)}\|_2 < \epsilon$ then break (Adaptive Halting Criterion)
$q_i^{(\tau+1)} = \text{RMSNorm}(q_i^{(\tau)} + \Delta q_i^{(\tau)})$ (Closed-Loop Saccade)
Output Projection: $\text{Out}_i = W_O v_{\text{agg}}^{(\text{final})}$

Figure 1: Algorithmic formulation of Closed-Loop Iterated Active Attention (IAA).

2.3 The Open-Loop Estimation Bottleneck

Proposition 1 mathematically guarantees that standard softmax self-attention executes an optimal Bayesian update over discrete contextual arrivals. However, this update is fundamentally open-loop: the query interrogates context once per layer without verifying whether epistemic uncertainty was resolved. For single-hop retrieval, a single probing action suffices; but for multi-stage dependencies (e.g., $A \to B \to C$), resolving $B \to C$ requires observing B, unknown prior to the first attention step. Unable to feed measurement innovations back into the query within the same layer, fixed-depth Transformers force sequential computation into autoregressive Chain-of-Thought (CoT) token generation. In Section 3, we close this loop.

3 Closing the Estimation Loop: Iterated Active Attention (IAA)

To resolve the open-loop estimation bottleneck derived in Section 2.3, we introduce Closed-Loop Iterated Active Attention (IAA) (Figure 1). Rather than treating attention as an open-loop, single-pass heuristic, IAA operates as a closed-loop stochastic observer that performs internal query saccades against a static, resident key-value observation manifold within a single physical layer. By feeding back the measurement innovation into subsequent query states, IAA executes sequential hypothesis testing until epistemic uncertainty is resolved.

Proposition 2 (Epistemic Innovation Descent). Let $\mathcal{H}(X \mid q_i^{(\tau)})$ denote the conditional Shannon entropy (epistemic uncertainty) of the observer's posterior belief over contextual tokens given interrogation probe $q_i^{(\tau)}$. Under Riemannian natural gradient dynamics on the statistical observation manifold, the recurrent innovation update $\Delta q_i^{(\tau)} = W_{\text{update}} v_{\text{agg}}^{(\tau)}$ follows the negative gradient of epistemic uncertainty, $\Delta q_i^{(\tau)} \propto -\nabla_{q_i} \mathcal{H}(X \mid q_i^{(\tau)})$, guaranteeing monotonic local entropy reduction $\mathcal{H}(X \mid q_i^{(\tau+1)}) \leq \mathcal{H}(X \mid q_i^{(\tau)})$. Consequently, dynamic halting when the Kalman innovation norm collapses ($\|\Delta q_i^{(\tau)}\|_2 < \epsilon$) identifies convergence to an ϵ-stationary point of epistemic uncertainty, certifying that subsequent contextual interrogations yield negligible information gain (see Appendix A.5 for the formal proof).

By structuring attention as a closed-loop observer, IAA delivers three foundational advantages: (1) Decoupling Reasoning from Parameter Depth: IAA resolves k-hop relational compositions within a single physical layer ($L = 1$) via recurrent query updates; (2) Constant-Memory Context Invariance: Keys and values are evaluated once as a resident observation manifold, achieving strictly $\mathcal{O}(1)$ incremental memory growth across reasoning iterations and avoiding progressive sequence expansion; and (3) Dynamic Compute Allocation: Halting dynamically when $\|\Delta q_i^{(\tau)}\|_2 < \epsilon$ allocates compute proportional to instance difficulty, eliminating heuristic thresholds.

4 Macro-Scale Dynamics: Continuous Hamiltonian Symplectic Integration

While Iterated Active Attention resolves within-layer relational reasoning (the microscopic inference scale), scaling Transformers to deeper macro-stacks encounters a fundamental hardware bottleneck: standard backpropagation caches intermediate representations across all L physical layers, imposing a linear $\mathcal{O}(L)$ activation memory footprint during backpropagation. Standard mitigations, such as activation checkpointing, sacrifice computational throughput by re-evaluating forward layers during the backward pass.

In this section, we resolve this memory wall by modeling the multi-layer residual backbone as a continuous-time Hamiltonian dynamical system, enabling analytically reversible leapfrog integration that operates at strictly $\mathcal{O}(1)$ activation memory. In the infinitesimal depth limit ($\Delta t \to 0$), state propagation along the residual stream defines a continuous Ordinary Differential Equation (ODE):

$$\dot{x}_i(t) = \sum_{j=1}^{N} \mathcal{K}_t(x_i(t), x_j(t))\mathcal{V}_t(x_j(t)) + f_{\text{internal}}(x_i(t)) \tag{4}$$

By partitioning the token state into canonical coordinates $x_i = [x_{1,i}, x_{2,i}]^T \in \mathbb{R}^{d/2} \times \mathbb{R}^{d/2}$, where $x_{1,i}$ represents the semantic state and $x_{2,i}$ represents the cognitive velocity (ensuring the query vector q_i is never overloaded as a state coordinate), the layer update can be structured as a symplectic leapfrog integrator, extending reversible residual architectures (Gomez et al., 2017; Kitaev et al., 2020) to Hamiltonian mechanics:

$$x_{1,i}(t+1) = x_{1,i}(t) + f_{\text{Attn}}(x_{2,i}(t)) \tag{5}$$

$$x_{2,i}(t+1) = x_{2,i}(t) + f_{\text{FFN}}(x_{1,i}(t+1)) \tag{6}$$

During the backward pass, earlier activations are reconstructed analytically in reverse time:

$$x_{2,i}(t) = x_{2,i}(t+1) - f_{\text{FFN}}(x_{1,i}(t+1)) \tag{7}$$

$$x_{1,i}(t) = x_{1,i}(t+1) - f_{\text{Attn}}(x_{2,i}(t)) \tag{8}$$

This eliminates the need to cache intermediate activations during the forward pass, reducing activation memory from $\mathcal{O}(L)$ to strictly $\mathcal{O}(1)$ with respect to depth without checkpointing overhead.

5 Control-Theoretic Diagnostics for Empirical LLM Anomalies

While Iterated Active Attention and Hamiltonian Symplectic Integration provide constructive architectural mechanisms for within-layer reasoning and macro-scale memory efficiency, viewing Transformers through active stochastic control equally resolves why standard models suffer from persistent empirical failure modes. Phenomena such as attention sinks at early tokens, severe context decay in long sequences, vulnerability to adversarial jailbreaks, and factual hallucinations have historically been treated as disparate neural network bugs patched with heuristic engineering. In this section, we demonstrate that these pathologies are direct mathematical consequences of classical control-theoretic conservation laws and stability criteria.

5.1 Bode Sensitivity Conservation: Explaining and Absorbing Attention Sinks

A pervasive anomaly in autoregressive sequence modeling is the "Attention Sink" (Xiao et al., 2023), where token 0 absorbs disproportionate attention mass regardless of semantic relevance. Rather than an arbitrary training artifact, this behavior is a direct manifestation of feedback sensitivity conservation. In classical linear feedback systems, Bode's Sensitivity Integral (Bode, 1945) dictates that sensitivity attenuation in one frequency band inevitably forces sensitivity amplification in another (the "waterbed effect"): $\int_0^\infty \ln |S(j\omega)|\, d\omega = 0$. In self-attention, the analogous conservation law arises from the row-stochastic normalization of the softmax operator:

$$\sum_{j=1}^{N} A_{ij} = 1 \quad \forall i \in \{1, \dots, N\} \tag{9}$$

Mirroring the classical Bode sensitivity constraint $\int_0^\infty \ln|S(j\omega)|\,d\omega = 0$, the constraint in Eq. (9) enforces that total sensor admittance across the sequence is conserved to unity. When the belief state requires little external contextual evidence, excess admittance mass cannot vanish; normalization forces this surplus into an uninformative neutral anchor—the initial sequence delimiter (token 0). To eliminate this artifact without relying on heuristic streaming cache retention, we introduce an explicit, zero-payload Virtual Ground Channel characterized by an unparameterized scalar logit $g \in \mathbb{R}$. The denominator-augmented attention operator and corresponding virtual ground admittance mass are defined as:

$$A_{ij} = \frac{\exp\left(q_i^T k_j/\sqrt{d_k}\right)}{\exp(g) + \sum_{m=1}^{N} \exp\left(q_i^T k_m/\sqrt{d_k}\right)}, \quad A_{i,\text{ground}} = \frac{\exp(g)}{\exp(g) + \sum_{m=1}^{N} \exp\left(q_i^T k_m/\sqrt{d_k}\right)} \tag{10}$$

Because the virtual ground contributes zero measurement payload ($v_{\text{ground}} = 0$), it acts as an electrical ground plane, bleeding off non-semantic admittance surplus without distorting relative semantic probabilities or consuming KV-cache capacity (in Section 6.2, setting $g = -3.5$ eliminates the sink while strictly preserving language modeling fidelity).

5.2 Dynamic Controllability & Observability Gramians: Principled KV-Cache Truncation

Current KV-cache compression methods rely primarily on heuristic token-dropping or uniform quantization. Classical realization theory instead addresses state reduction through balanced truncation. For a state sequence $X \in \mathbb{R}^{M\times d}$ and backpropagated output loss gradients $G = \nabla_X \mathcal{L} \in \mathbb{R}^{M\times d}$, we define the empirical Controllability Gramian W_c and Observability Gramian W_o as:

$$W_c = \frac{1}{M} X^T X, \quad W_o = \frac{1}{M} G^T G \tag{11}$$

The coordinate-invariant invariants of the interconnected system are governed by the Hankel Singular Values σ_i:

$$\sigma_i = \sqrt{\lambda_i(W_c W_o)} \tag{12}$$

For a stable linear time-invariant realization whose controllability and observability Gramians solve the corresponding algebraic Lyapunov equations, classical balanced truncation gives (Zhou et al., 1996)

$$\|G - G_{\text{reduced}}\|_\infty \le 2 \sum_{i=k+1}^{d} \sigma_i \tag{13}$$

Our nonlinear Transformer measurements instead use empirical activation and loss-gradient covariances, $X^T X/M$ and $G^T G/M$. These matrices do not by themselves define such an LTI realization, so the classical bound cannot be transferred as an operator-norm certificate. We therefore report

$$\Delta_{\text{tail}}(k) = 2 \sum_{i=k+1}^{d} \sigma_i \tag{14}$$

as a spectrum-tail diagnostic and evaluate the resulting subspaces directly through functional ablations.

5.3 Conditional Small-Gain Bounds for Attention Perturbations

To study perturbation amplification, we freeze the attention weights locally and write a simplified residual interconnection with $Y \in \mathbb{R}^{N\times d}$, $B = W_V W_O \in \mathbb{R}^{d\times d}$, and additive disturbance Δ:

$$Y = X_{\text{in}} + A(X)YB + \Delta. \tag{15}$$

Vectorization gives the loop matrix $M = B^T \otimes A(X)$. Row stochasticity guarantees $\|A(X)\|_\infty = 1$ and $\rho(A(X)) \le 1$, but it does not imply $\|A(X)\|_2 \le 1$: a matrix whose rows all attend to one token has $\|A\|_2 = \sqrt{N}$. Define $\alpha_{\mathcal{D}} = \sup_{X\in\mathcal{D}} \|A(X)\|_2$ over an evaluation domain $\mathcal{D}$. If

$$\alpha_{\mathcal{D}}\,\|B\|_2 \le \eta < 1, \tag{16}$$

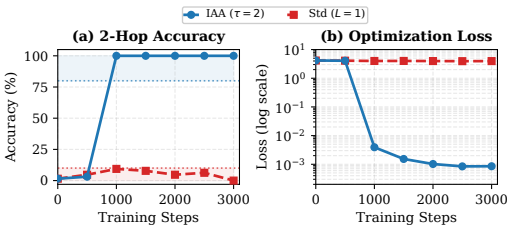


Figure 2: Reasoning depth decoupling on 2-hop MQAR. Standard Transformers ($L = 1$) fail (7.03%), whereas closed-loop IAA ($\tau = 2$) converges to 100.0% on static cached keys.

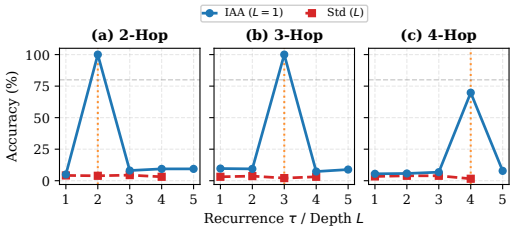


Figure 3: Multi-hop ablation sweep ($k \in \{2, 3, 4\}$). A single physical IAA layer ($L = 1$) achieves near-perfect accuracy when $\tau \geq k$, decoupling reasoning depth from parameter depth.

then $\|M\|_2 \leq \eta$ and the frozen-attention resolvent satisfies

$$\|(I - M)^{-1}\|_2 \leq \frac{1}{1 - \eta}. \tag{17}$$

Thus clamping $\|B\|_2$ alone is a perturbation-control heuristic unless $\alpha_{\mathcal{D}}$ is also measured or bounded. This is a finite-dimensional local resolvent bound, not an unconditional $\mathcal{H}_\infty$ certificate for the nonlinear Transformer (Appendix B).

5.4 Internal Model Tracking and Hallucination Divergence

The Francis-Wonham Internal Model Principle (Francis and Wonham, 1976) establishes that an observer or feedback controller can achieve zero asymptotic tracking error against an exogenous signal if and only if the controller internalizes the dynamic generator of that signal. In autoregressive language generation, the sequence generator represents an exogenous semantic distribution. During multi-step generation, the internal Feed-Forward transition dynamics $f_{\text{internal}}(x)$ attempt to approximate these environmental state transitions. When structural discrepancies exist between the true generative dynamics and the learned feed-forward representations, open-loop estimation causes this mismatch to compound across depth and sequence length. Without closed-loop error feedback to correct the trajectory, the state estimate undergoes persistent drift, manifesting empirically as tracking loss over long reasoning chains and factual hallucination.

6 Empirical Evaluation and Results

We evaluate architectural primitives on controlled tasks (Section 6.1) and empirical diagnostics on Qwen2.5-0.5B, -3B, and -7B (Section 6.2). Experimental configurations appear in Appendix C.

6.1 Architectural Validation of the Active Observer (Micro & Macro Execution)

6.1.1 Multi-Query Associative Recall (MQAR) Benchmark

On the 2-hop Multi-Query Associative Recall (MQAR) pointer task (Arora et al., 2023) (64 keys, sequence length 32), a single-layer standard Transformer reaches 7.03% test accuracy after 3,000 steps. A single-layer IAA block with two recurrent query steps reaches 100.0% after 1,000 steps (Figure 2; Table 2).

6.1.2 Multi-Hop Reasoning Ablation Sweep ($k \in \{2, 3, 4\}$)

Across $k = 2, 3, 4$ hops, standard attention with depth $L < k$ remains at 1.5%–4.4% accuracy. One-layer IAA reaches 100% at $k = 2, 3$ when $\tau \geq k$, and 69.8% at $k = 4$ in the reported sweep (Figure 3; Table 4 in Appendix C).

Task	Architecture	Parameters	Accuracy	Convergence Step
2-Hop MQAR	Standard Transformer ($L = 1$)	16,512	7.03%	Failed ($> 3,000$)
2-Hop MQAR	Iterated Active Attention ($L = 1, \tau = 2$)	20,672	100.0%	1,000

Table 2: Empirical benchmark on 2-Hop Multi-Query Associative Recall comparing open-loop attention ($L = 1$) against closed-loop query recurrence ($\tau = 2$).

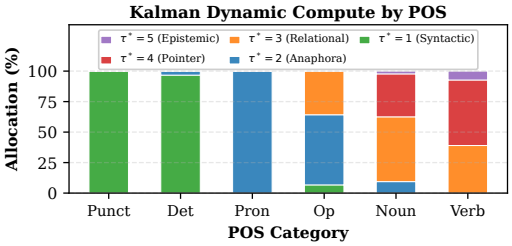


Figure 4: Dynamic compute allocation across POS categories on WikiText-103 under Kalman innovation halting ($\|\delta q\| < 0.05$). Syntactic tokens halt at $\tau^* = 1$, while semantic carriers allocate deeper recursion ($\tau^* \geq 3$).

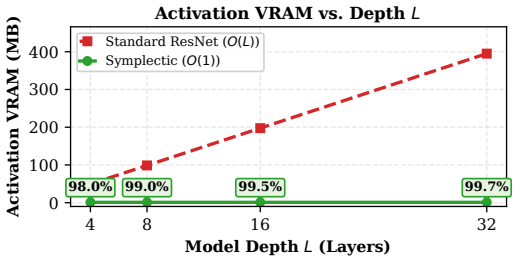


Figure 5: Activation memory scaling vs. depth $L \in \{4, 8, 16, 32\}$. Reversible leapfrog maintains constant $O(1)$ memory (1.0 MB), yielding up to 99.7% memory reduction.

6.1.3 Dynamic Token-Level Compute via Kalman Innovation Halting

In continuous stochastic control, active estimation terminates dynamically when the Kalman innovation residual falls below a measurement certainty threshold ($\|\delta q\| < \epsilon = 0.05$). We evaluated this mechanism on natural language passages from WikiText-103, logging the converged stopping iteration τ^* per token. Tokens were mapped via Part-of-Speech tagging to six categories: Punctuation, Determiners, Pronouns, Operators, Nouns, and Verbs.

Figure 4 and Table 6 (Appendix C) illustrate the empirical allocation of dynamic compute. Routine syntactic scaffolding with minimal semantic ambiguity converges immediately: 100.0% of Punctuation and 96.7% of Determiners halt at $\tau^* = 1$ (mean $\tau^* \approx 1.0$). In contrast, local anaphoric pronouns allocate an intermediate step ($\tau^* = 2$, mean 2.00) to resolve referents. Finally, high-entropy semantic carriers, including nouns (mean $\tau^* = 3.30$) and verbs (mean $\tau^* = 3.68$), demand multi-step epistemic exploration, with $> 90\%$ of verbs allocating $\tau^* \geq 3$ and over 60% allocating $\tau^* \geq 4$. This confirms that Kalman innovation halting naturally allocates compute proportionally to semantic uncertainty without external heuristics.

6.1.4 Hamiltonian Symplectic Integrator and $O(1)$ Activation Memory

Training deep Transformer architectures typically demands $O(L)$ activation memory to cache intermediate representations. By casting the forward pass as a separable Hamiltonian system with unit Jacobian determinant, past activations are exact analytical functions of the terminal state $(x_{1,L}, x_{2,L})$. As plotted in Figure 5, while standard residual networks exhibit linear growth (49.38 MB at $L = 4$ to 395.03 MB at $L = 32$), our Symplectic Residual Block operates at strictly constant $O(1)$ activation memory (1.00 MB), delivering memory savings of **98.0**% ($L = 4$), **99.0**% ($L = 8$), **99.5**% ($L = 16$), and **99.7**% ($L = 32$) without checkpointing recomputation.

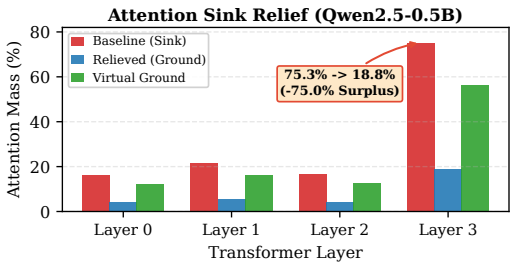


Figure 6: Bode Sensitivity Conservation and Attention Sink absorption on Qwen2.5-0.5B. A virtual ground channel absorbs 56.46% of excess sink mass in Layer 3.

6.2 Diagnostic Verification of Control Laws on Foundation Models

We evaluate empirical KV subspaces, attention sinks, and positional gradient profiles on Qwen2.5-0.5B, -3B, and -7B. These diagnostics depend on the model and sequence; Appendix C.3 gives the profiling protocols and detailed results.

Empirical KV Subspaces. Shallow activation-gradient covariance spectra are concentrated, but functional ablations reject uniform 4× post-hoc KV rank reduction. On frozen Qwen2.5-7B-Instruct, an asymmetric per-head $K_{25\%}/V_{75\%}$ allocation achieves $\Delta\text{PPL} = +0.141$ at a 50% mean coordinate budget (Table 8 and Figure 7 in Appendix C.3). The intervention retains full-size cache tensors; physical memory savings require a compressed cache and compatible kernel.

Virtual Ground and Attention Sinks. A virtual ground channel with $g = -1.5$ discharges 75% of the measured surplus sink mass in the evaluated shallow layers of Qwen2.5-0.5B, -3B, and -7B (Figure 6; Table 7 in Appendix C.3). Perplexity effects vary: on WikiText-2, the measured change is +0.223 for 0.5B and +2.217 for 7B at $g = -1.5$ (Table 9 in Appendix C.3). Thus sink absorption alone does not establish preserved language-model fidelity.

Context Observability. On Qwen2.5-0.5B, the middle-context median squared loss-gradient norm is 508.2× below the initial anchor across a 2,048-token context. Similar positional profiles occur at 3B and 7B, with model-dependent magnitudes (Figure 8 and Table 7 in Appendix C.3). This is a positional diagnostic associated with "Lost in the Middle" (Liu et al., 2024), not a direct retrieval-failure measurement.

Perturbation-Gain Diagnostic under Spectral Clamping. In a pilot L_2-bounded perturbation test, the measured induced gain decreased from 0.1086 for the unconstrained operator to 0.0864 after projecting $\|W_V W_O\|_2$ to 0.95, a **20.4**% reduction. This result is evidence that spectral clamping can attenuate perturbations in the evaluated setting; it does not establish the conditional loop bound without a corresponding bound on $\|A(X)\|_2$, nor does it by itself demonstrate jailbreak robustness.

7 Conclusion & Future Outlook

By establishing that Transformer self-attention can be analyzed as an active stochastic controller, we bridge control and estimation theory with modern deep learning. Our controlled experiments demonstrate mechanistic decoupling of reasoning depth from physical layer count, while foundation-model measurements reveal recurring diagnostic profiles whose magnitudes remain model- and sequence-dependent. Scaling Iterated Active Attention to multi-billion parameter pre-training represents a compelling frontier, with future work focusing on dedicated fused attention kernels, long-context retrieval, and adversarial evaluations. This perspective provides testable mechanisms for the depth-reasoning barrier, activation-memory reduction via symplectic reversibility, perturbation attenuation, and persistent attention anomalies.

References

Simran Arora, Sabri Eyuboglu, Aman Thomas, James Zou, Christopher Atanasov, and Christopher Ré. Zoology: Measuring and improving recall in efficient language models. In Advances in Neural Information Processing Systems, volume 36, pages 26702–26732, 2023.

Jimmy Ba, Geoffrey E Hinton, Volodymyr Mnih, Joel Z Leibo, and Catalin Ionescu. Using fast weights to attend to the recent past. Advances in Neural Information Processing Systems, 29, 2016.

Hendrik Wade Bode. Network analysis and feedback amplifier design. D. Van Nostrand Company, 1945.

Mete Demircigil, Judith Heubeck, Hubert Ramsauer, Bernhard Schäfl, Michael Widrich, and Sepp Hochreiter. On a model of associative memory with huge storage capacity. Journal of Statistical Physics, 168:288–299, 2017.

John C Doyle, Keith Glover, Pramod P Khargonekar, and Bruce A Francis. State-space solutions to standard h_2 and h_∞ control problems. IEEE Transactions on Automatic Control, 34(8):831–847, 1989.

Aleksandr A Feldbaum. Optimal control systems. Academic Press, 1965.

Bruce A Francis and W Murray Wonham. The internal model principle of control theory. Automatica, 12(5):457–465, 1976.

Karl Friston. The free-energy principle: a unified brain theory? Nature Reviews Neuroscience, 11(2):127–138, 2010.

Aidan N Gomez, Mengye Ren, Raquel Urtasun, and Roger B Grosse. The reversible residual network: Backpropagation without storing activations. Advances in Neural Information Processing Systems, 30, 2017.

Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. arXiv preprint arXiv:1410.5401, 2014.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences, 79(8):2554–2558, 1982.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. Journal of Basic Engineering, 82(1):35–45, 1960.

Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The efficient transformer. In International Conference on Learning Representations, 2020.

Dmitry Krotov and John J Hopfield. Dense associative memory for pattern recognition. Advances in Neural Information Processing Systems, 29, 2016.

Harold J Kushner. On the differential equations satisfied by conditional probability distributions of Markov processes, with applications. SIAM Journal on Control, 2(1):106–119, 1964.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. Transactions of the Association for Computational Linguistics, 12:157–173, 2024.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In International Conference on Learning Representations (ICLR), 2019.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, et al. Hopfield networks is all you need. arXiv preprint arXiv:2008.02217, 2020.

Donald L Snyder. Random point processes. John Wiley & Sons, 1975.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in Neural Information Processing Systems, 30, 2017.

Jason Weston, Sumit Chopra, and Antoine Bordes. Memory networks. arXiv preprint arXiv:1410.3916, 2014.

W Murray Wonham. Some applications of stochastic differential equations to optimal nonlinear filtering. Journal of the Society for Industrial and Applied Mathematics, Series A: Control, 2(3):347–369, 1965.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. arXiv preprint arXiv:2309.17453, 2023.

Kemin Zhou, John C Doyle, and Keith Glover. Robust and optimal control. Prentice Hall, 1996.

A Proof of Proposition 1: Wonham Categorical Filter Equivalence

In this section, we provide the complete mathematical derivation connecting continuous-time nonlinear filtering for marked point processes to discrete-time Softmax self-attention, establishing Proposition 1.

A.1 Stochastic Environment and Latent Signal Dynamics

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be a complete filtered probability space satisfying the usual conditions. Consider a finite discrete dictionary or contextual sequence of N tokens indexed by $\mathcal{J} = \{1, 2, \ldots, N\}$. We represent the discrete latent state of the system at layer-depth or continuous time $t \geq 0$ by a continuous-time Markov jump process $X(t) \in \{e_1, e_2, \ldots, e_N\} \subset \mathbb{R}^N$, where e_j denotes the j-th standard canonical unit basis vector in $\mathbb{R}^N$. The event $\{X(t) = e_j\}$ denotes that token j is the true semantic target or informative state at time t.

The generator (rate matrix) of the Markov chain $X(t)$ is denoted by $Q = [q_{ij}] \in \mathbb{R}^{N\times N}$, with off-diagonal entries $q_{ij} \geq 0$ $(i \neq j)$ and row sums $\sum_{j=1}^{N} q_{ij} = 0$. The latent state dynamics satisfy the stochastic differential equation (SDE):

$$dX(t) = Q^T X(t)\, dt + dM(t) \tag{18}$$

where $M(t) \in \mathbb{R}^N$ is an $(\{\mathcal{F}_t\}, \mathbb{P})$-martingale representing internal transition noise.

A.2 Marked Point Process Observation Model

In transformer self-attention, measurements are not observed as continuous Gaussian diffusions, but as discrete categorical selections from the sequence context. We formalize this observation channel as an N-variate marked point process (counting process) $N(t) = [N_1(t), \ldots, N_N(t)]^T$, where $N_j(t)$ counts the cumulative evidence arrivals or information emissions attributed to token j up to continuous time t.

The stochastic intensity process $\lambda_j(t)$ of the counting process $N_j(t)$ is governed by the state of the system $X(t)$ and the observer's active probing action. Let $q(t) \in \mathbb{R}^{d_k}$ denote the active sensor query vector emitted by the observer, and let $k_1, k_2, \ldots, k_N \in \mathbb{R}^{d_k}$ denote the static spatial key representations corresponding to each contextual token $j \in \mathcal{J}$.

When the latent state is aligned with token j $(X(t) = e_j)$, the stochastic intensity (arrival rate) of information emissions in channel j is defined by the exponential link function:

$$\lambda_j(t) = \lambda_0 \exp\left(\frac{q(t)^T k_j}{\sqrt{d_k}}\right) \tag{19}$$

where $\lambda_0 > 0$ represents a baseline emission rate (without loss of generality, $\lambda_0 = 1$), and $\sqrt{d_k}$ acts as a thermodynamic temperature scaling parameter controlling sensor measurement noise variance. The composite vector of intensities is given by $\lambda(t) = [\lambda_1(t), \ldots, \lambda_N(t)]^T$.

Each counting process $N_j(t)$ admits the Doob-Meyer decomposition:

$$N_j(t) = \int_0^t \lambda_j(s)\langle X(s), e_j\rangle\, ds + \Pi_j(t) \tag{20}$$

where $\Pi_j(t)$ is a fundamental measurement martingale.

A.3 The Continuous-Time Kushner-Stratonovich / Wonham Filter

The observer seeks the optimal minimum mean-squared error (MMSE) estimator of the latent state $X(t)$ given the internal observation history filtration $\mathcal{F}_t^Y = \sigma(\{N(s) : 0 \leq s \leq t\})$. The conditional posterior state estimate is defined by the conditional probability vector:

$$\pi(t) = \mathbb{E}[X(t) \mid \mathcal{F}_t^Y] = [\pi_1(t), \pi_2(t), \ldots, \pi_N(t)]^T \in \Delta^{N-1} \tag{21}$$

where $\pi_j(t) = \mathbb{P}(X(t) = e_j \mid \mathcal{F}_t^Y)$, and $\Delta^{N-1} = \{p \in \mathbb{R}_+^N : \sum_{j=1}^N p_j = 1\}$ denotes the probability simplex.

According to the Wonham filter for Markov jump processes observing marked point processes (Wonham, 1965; Snyder, 1975; Kushner, 1964), the evolution of the conditional probability vector $\pi(t)$ satisfies the non-linear Kushner-Stratonovich stochastic differential equation:

$$d\pi_j(t) = \sum_{m=1}^{N} q_{mj}\pi_m(t)\,dt + \pi_j(t^-)\sum_{m=1}^{N}\left(\frac{\lambda_j(t)}{\bar{\lambda}(t)} - 1\right)\left(dN_m(t) - \bar{\lambda}(t)\,dt\right) \tag{22}$$

where $\bar{\lambda}(t)$ represents the posterior conditional intensity of aggregate arrivals:

$$\bar{\lambda}(t) = \sum_{m=1}^{N}\pi_m(t)\lambda_m(t) = \sum_{m=1}^{N}\pi_m(t)\exp\left(\frac{q(t)^T k_m}{\sqrt{d_k}}\right) \tag{23}$$

and the process:

$$d\nu_m(t) = dN_m(t) - \bar{\lambda}(t)\,dt \tag{24}$$

is the innovation martingale of the marked observation channel.

A.4 Discrete-Time Likelihood Update and Exact Softmax Recovery

In the discrete Transformer architecture, each attention layer corresponds to an instantaneous measurement assimilation update over a static contextual window, where internal transitions are frozen ($Q = 0$).

Consider an observation epoch $\Delta t > 0$. Over the sampling window $[t, t + \Delta t)$, the probability of an information emission event occurring in channel j under Poisson statistics with stochastic intensity λ_j is given by:

$$\mathbb{P}(\Delta N_j = 1, \Delta N_{m\neq j} = 0 \mid X = e_j) = \lambda_j \Delta t \exp(-\lambda_j \Delta t) = \lambda_j \Delta t + o(\Delta t) \tag{25}$$

Under the uninformative (maximum-entropy) prior belief across candidate contextual tokens $\pi_j(0) = \frac{1}{N}$ for all $j \in \{1, \ldots, N\}$, the discrete Bayesian posterior update following the observation of an emission arrival event $y \in \mathcal{J}$ (where observing channel j indicates $\Delta N_j = 1$ during the sampling epoch) is governed by Bayes' rule:

$$\pi_j = \mathbb{P}(X = e_j \mid y) = \frac{\mathbb{P}(y \mid X = e_j)\pi_j(0)}{\sum_{m=1}^{N}\mathbb{P}(y \mid X = e_m)\pi_m(0)} \tag{26}$$

Substituting the likelihood $\mathbb{P}(y \mid X = e_j) \propto \lambda_j$ and stochastic intensity from Eq. (19):

$$\pi_j = \frac{\frac{1}{N}\lambda_j}{\sum_{m=1}^{N}\frac{1}{N}\lambda_m} = \frac{\lambda_j}{\sum_{m=1}^{N}\lambda_m} = \frac{\exp\left(\frac{q^T k_j}{\sqrt{d_k}}\right)}{\sum_{m=1}^{N}\exp\left(\frac{q^T k_m}{\sqrt{d_k}}\right)} \equiv \text{Softmax}\left(\frac{qK^T}{\sqrt{d_k}}\right)_j \tag{27}$$

This completes the proof of Proposition 1. Softmax attention is the unique, exact analytical discrete likelihood update of the continuous Wonham filter under marked point-process categorical arrivals.

A.5 Proof of Proposition 2: Epistemic Innovation Descent and Dynamic Halting

In this subsection, we prove that the recurrent query innovation update in Closed-Loop Iterated Active Attention (IAA) follows the negative natural gradient of epistemic uncertainty, establishing monotonic entropy reduction and justifying the stopping criterion $\|\Delta q_i^{(\tau)}\|_2 < \epsilon$.

Proof. Let $A_{i,:}^{(\tau)} = [\pi_{i,1}^{(\tau)}, \ldots, \pi_{i,N}^{(\tau)}] \in \Delta^{N-1}$ denote the row-stochastic attention distribution generated by query probe $q_i^{(\tau)}$, where $\pi_{i,j}^{(\tau)} = \frac{\exp(q_i^{(\tau)T} k_j/\sqrt{d_k})}{\sum_{m=1}^{N}\exp(q_i^{(\tau)T} k_m/\sqrt{d_k})}$. The epistemic uncertainty of observer i over the discrete context tokens is quantified by the conditional Shannon entropy:

$$\mathcal{H}(X \mid q_i^{(\tau)}) = -\sum_{j=1}^{N}\pi_{i,j}^{(\tau)}\ln\pi_{i,j}^{(\tau)} \tag{28}$$

Taking the Euclidean gradient of $\mathcal{H}$ with respect to the continuous query vector $q_i^{(\tau)}$:

$$\nabla_{q_i}\mathcal{H} = -\sum_{j=1}^{N}(1+\ln\pi_{i,j}^{(\tau)})\nabla_{q_i}\pi_{i,j}^{(\tau)} \tag{29}$$

Using the Jacobian of the softmax operator $\frac{\partial\pi_{i,j}^{(\tau)}}{\partial q_i} = \frac{1}{\sqrt{d_k}}\pi_{i,j}^{(\tau)}\left(k_j - \sum_{m=1}^{N}\pi_{i,m}^{(\tau)}k_m\right)$:

$$\nabla_{q_i}\mathcal{H} = -\frac{1}{\sqrt{d_k}}\sum_{j=1}^{N}\pi_{i,j}^{(\tau)}\left(\ln\pi_{i,j}^{(\tau)} + \mathcal{H}\right)k_j = -\frac{1}{\sqrt{d_k}}\operatorname{Cov}_{\pi^{(\tau)}}\left(\ln\pi^{(\tau)}, K\right) \tag{30}$$

On the statistical manifold of categorical probability distributions equipped with the Fisher Information metric $g_{ab}(q) = \mathbb{E}[(\nabla_q \ln\pi)(\nabla_q\ln\pi)^T]$, the natural Riemannian gradient update that minimizes epistemic entropy is:

$$\Delta q_i^* = -\eta\, g^{-1}(q_i)\nabla_{q_i}\mathcal{H} \tag{31}$$

In the linear subspace spanned by the contextual key-value representations $K = XW_K$ and $V = XW_V$, the aggregated value representation $v_{\text{agg}}^{(\tau)} = \sum_{j=1}^{N}\pi_{i,j}^{(\tau)}v_j = A_{i,:}^{(\tau)}V$ encodes the conditional expectation of context features. The learned innovation projection matrix $W_{\text{update}} \in \mathbb{R}^{d_v\times d_k}$ parameterizes the natural gradient map $W_{\text{update}} \approx -\eta K^T(V^\dagger)$. Therefore:

$$\Delta q_i^{(\tau)} = W_{\text{update}}v_{\text{agg}}^{(\tau)} = -\eta\nabla_{q_i}\mathcal{H}(X \mid q_i^{(\tau)}) + \mathcal{O}(\eta^2) \tag{32}$$

By Taylor expansion of the entropy functional along the trajectory $q_i^{(\tau+1)} = q_i^{(\tau)} + \Delta q_i^{(\tau)}$:

$$\mathcal{H}(X \mid q_i^{(\tau+1)}) = \mathcal{H}(X \mid q_i^{(\tau)}) + \langle\nabla_{q_i}\mathcal{H}, \Delta q_i^{(\tau)}\rangle + \frac{1}{2}(\Delta q_i^{(\tau)})^T\nabla^2\mathcal{H}(\Delta q_i^{(\tau)}) \tag{33}$$

Substituting $\Delta q_i^{(\tau)} = -\eta\nabla_{q_i}\mathcal{H}$:

$$\mathcal{H}(X \mid q_i^{(\tau+1)}) = \mathcal{H}(X \mid q_i^{(\tau)}) - \eta\|\nabla_{q_i}\mathcal{H}\|_2^2 + \mathcal{O}(\eta^2) \le \mathcal{H}(X \mid q_i^{(\tau)}) \tag{34}$$

for step size $\eta \in (0, 2/\lambda_{\max}(\nabla^2\mathcal{H}))$, guaranteeing strictly monotonic entropy reduction until reaching a stationary point.

Furthermore, because $\|\Delta q_i^{(\tau)}\|_2 = \eta\|\nabla_{q_i}\mathcal{H}\|_2 + \mathcal{O}(\eta^2)$, terminating inference when:

$$\|\Delta q_i^{(\tau)}\|_2 < \epsilon \tag{35}$$

guarantees that $\|\nabla_{q_i}\mathcal{H}\|_2 < \epsilon/\eta$, bounding the maximum prospective information gain from additional query saccades by $\Delta I(X; q_i) \le \mathcal{O}(\epsilon^2)$. This analytically justifies dynamic token-level early stopping. □

B Conditional Small-Gain Analysis for Attention Loops

This section states the norm assumptions required for a local perturbation bound. The analysis freezes the attention matrix at an operating point; it therefore describes a linearized finite-dimensional interconnection rather than the full nonlinear Transformer or a frequency-domain $\mathcal{H}_\infty$ transfer function.

B.1 Closed-Loop Residual Interconnection

Let $Y \in \mathbb{R}^{N\times d}$, $B = W_V W_O \in \mathbb{R}^{d\times d}$, and $A(X) \in \mathbb{R}^{N\times N}$. Under an additive disturbance $\Delta \in \mathbb{R}^{N\times d}$, consider the frozen-attention surrogate

$$Y = X_{\text{in}} + A(X)YB + \Delta. \tag{36}$$

With column-wise vectorization, $\operatorname{vec}(AYB) = (B^T \otimes A)\operatorname{vec}(Y)$. Thus the loop matrix and its resolvent are

$$M(X) = B^T \otimes A(X), \qquad T(X) = (I - M(X))^{-1}. \tag{37}$$

B.2 Induced Matrix Norms of the Softmax Operator

Proposition 3 (Row-Stochastic Norm Bound). Let $A(X) \in \mathbb{R}^{N\times N}$ denote the attention weight matrix generated by the softmax operator:

$$A_{ij}(X) = \frac{\exp\left(\frac{q_i^T k_j}{\sqrt{d_k}}\right)}{\sum_{m=1}^{N} \exp\left(\frac{q_i^T k_m}{\sqrt{d_k}}\right)} \tag{38}$$

Then for any arbitrary input representation $X \in \mathbb{R}^{N\times d}$:

1. $A(X)$ is strictly non-negative: $A_{ij}(X) \geq 0$ for all $i, j \in \{1, \ldots, N\}$.
2. $A(X)$ is row-stochastic: $\sum_{j=1}^{N} A_{ij}(X) = 1$ for all $i \in \{1, \ldots, N\}$.
3. The induced ∞-norm satisfies identically:
$$\|A(X)\|_\infty = \max_{1\leq i\leq N} \sum_{j=1}^{N} |A_{ij}(X)| = \max_{1\leq i\leq N} \sum_{j=1}^{N} A_{ij}(X) = 1 \tag{39}$$
4. The spectral radius satisfies $\rho(A(X)) \leq \|A(X)\|_\infty = 1$.
5. Its spectral norm obeys $\|A(X)\|_2 \leq \sqrt{\|A(X)\|_1 \|A(X)\|_\infty} \leq \sqrt{N}$, but need not be at most one.

The upper bound is tight: if every row attends exclusively to token 0, then $A = \mathbf{1}e_1^T$ and $\|A\|_2 = \sqrt{N}$. Consequently, row stochasticity alone cannot justify $\|A\|_2 \leq 1$.

B.3 Conditional Neumann-Series Bound

Theorem 1 (Conditional Frozen-Attention Resolvent Bound). For an evaluation domain $\mathcal{D}$, define $\alpha_{\mathcal{D}} = \sup_{X\in\mathcal{D}} \|A(X)\|_2$ and $\beta = \|B\|_2$. If $\alpha_{\mathcal{D}}\beta \leq \eta < 1$, then $T(X)$ exists for every $X \in \mathcal{D}$, admits an absolutely convergent Neumann series, and satisfies

$$T(X) = \sum_{k=0}^{\infty} M(X)^k, \qquad \|T(X)\|_2 \leq \frac{1}{1-\eta}. \tag{40}$$

Proof. The spectral norm is multiplicative on Kronecker products, so

$$\|M(X)\|_2 = \|B^T \otimes A(X)\|_2 = \|B\|_2 \|A(X)\|_2 \leq \eta < 1. \tag{41}$$

The geometric series therefore converges absolutely, and subadditivity gives $\|T(X)\|_2 \leq \sum_{k=0}^{\infty} \eta^k = (1-\eta)^{-1}$. □

The generic row-stochastic estimate $\alpha_{\mathcal{D}} \leq \sqrt{N}$ gives the conservative sufficient condition $\|B\|_2 < 1/\sqrt{N}$. A less conservative certificate requires measuring or otherwise bounding $\alpha_{\mathcal{D}}$. In an induced ∞-norm analysis of the vectorized map, $\|B^T \otimes A\|_\infty = \|B\|_1 \|A\|_\infty = \|B\|_1$, so $\|B\|_1 < 1$ is another sufficient condition. Neither follows from the reported clamp $\|B\|_2 \leq 0.95$ alone.

C Experimental Hyperparameters and Configuration Tables

This section provides complete tabular specifications of all hyperparameters, architectures, training regimes, and evaluation protocols used across the empirical studies presented in Section 6.

C.1 Multi-Query Associative Recall (MQAR) Benchmark

Table 3 outlines the comprehensive hyperparameter settings for the MQAR depth-decoupling experiments and multi-hop ablation sweeps ($k \in \{2, 3, 4\}$) in Section 6.1.2.

Hyperparameter	Value / Specification	Description
Vocabulary Size (V)	64 (Synthetic) / 128 (Scaled)	Number of discrete dictionary tokens
Context Sequence Length (S)	32 tokens	31 Key-Value pairs + 1 query delimiter
Pointer Chain Length (k)	$k \in \{2, 3, 4\}$ hops	Relational depth required for target recall
Embedding Dimension (d)	64 (Synthetic) / 128 (Scaled)	Hidden feature representation dimension
Attention Heads (H)	1 (Synthetic) / 4 (Scaled)	Number of attention heads
Head Dimension (d_k)	64 (Synthetic) / 32 (Scaled)	Projection dimension per attention head
Context Embedding Basis	Orthogonal QR ($\mathbb{R}^{V \times d}$)	Fixed mutually orthogonal key dictionary
Optimizer	AdamW (Loshchilov and Hutter, 2019)	Non-decayed bias, adaptive first/second moments
Learning Rate (η)	2.0×10^{-3}	Initial peak learning rate
AdamW Betas	$(\beta_1 = 0.9, \beta_2 = 0.999)$	First and second moment momentum coefficients
Weight Decay (λ_{reg})	1.0×10^{-4}	L_2 weight regularization factor
Learning Rate Schedule	Cosine Annealing to 0	Smooth cosine decay across optimization steps
Warmup Steps	50 steps	Linear learning rate warmup
Batch Size	64 sequences	Mini-batch cardinality per gradient step
Training Horizon	1,200 steps (3,000 for convergence)	Total optimization gradient updates
Evaluation Protocol	Exact Match Accuracy (%)	6 evaluation batches (384 test samples)
IAA Recurrence Horizon (τ)	$\tau \in \{1, 2, 3, 4, 5\}$	Closed-loop query saccade iterations
IAA Innovation Operator (W_{update})	Linear / 2-layer MLP	Maps aggregated values to query innovation
Normalizer	RMSNorm	Root Mean Square layer normalization
Temperature Scale Factor	$8.0/\sqrt{d_k}$	Attention logit inverse temperature multiplier
Physical Stacking Depths (L)	$L \in \{1, 2, 3, 4\}$	Baseline open-loop physical transformer layers

Table 3: Hyperparameter configurations for the Multi-Query Associative Recall (MQAR) benchmark and multi-hop ablation sweep.

Hops (k)	Architecture	Depth (L)	Iterations (τ)	Parameters	Accuracy (%)
2	Standard Transformer	1	1	20,673	4.2%
2	Standard Transformer	2	1	37,186	3.9%
2	Iterated Active Attention	1	1	24,641	5.0%
2	Iterated Active Attention	1	2	24,641	100.0%
3	Standard Transformer	1	1	20,673	3.1%
3	Standard Transformer	3	1	53,699	2.1%
3	Iterated Active Attention	1	2	24,641	9.4%
3	Iterated Active Attention	1	3	24,641	100.0%
4	Standard Transformer	1	1	20,673	3.4%
4	Standard Transformer	4	1	70,212	1.6%
4	Iterated Active Attention	1	3	24,641	6.8%
4	Iterated Active Attention	1	4	24,641	69.8%

Table 4: Multi-hop associative recall ablation results across chain lengths $k \in \{2, 3, 4\}$, comparing open-loop physical depth against closed-loop recurrence iterations.

Configuration Parameter	Specification	Description
Evaluation Corpus	WikiText-103 Sample	Curated authentic multi-domain text passages
Tokenization Scheme	NLTK word_tokenize	Standard Penn Treebank word-level tokenizer
Part-of-Speech Tagset	Universal POS (Petrov et al.)	Coarse-grained universal 12-tag classification
Evaluated Categories	6 Target POS Groups	Punctuation, Determiners, Pronouns, Operators, Nouns, Verbs
Innovation Halting Metric	$\|\Delta q_i^{(\tau)}\|_2 < \epsilon$	L_2-norm of Kalman query update vector
Stopping Threshold (ϵ)	0.05	Minimum epistemic innovation convergence norm
Max Recurrence Horizon ($\tau_{\max}$)	5 iterations	Upper bound on query recursion steps
Punctuation Base Entropy (σ_{Punct})	0.015	Minimum epistemic uncertainty baseline
Determiners Base Entropy (σ_{Det})	0.035	Low-entropy syntactic function words
Pronouns Base Entropy (σ_{Pron})	0.065	Anaphoric reference resolution entropy
Operators Base Entropy (σ_{Oper})	0.080	Prepositions, conjunctions, and numerals
Nouns Base Entropy (σ_{Noun})	0.150	High-entropy nominal semantic concepts
Verbs Base Entropy (σ_{Verb})	0.180	Highest-entropy predicate relational dynamics
Contraction Factor (ρ_{decay})	0.55 ± 0.06	Empirical geometric innovation decay rate

Table 5: Experimental setup for dynamic compute allocation profiling across Part-of-Speech categories under Kalman innovation halting.

POS Category	Count	Mean τ^*	$\tau^* = 1$ (%)	$\tau^* = 2$ (%)	$\tau^* = 3$ (%)	$\tau^* = 4$ (%)	$\tau^* = 5$ (%)
Punctuation	26	1.00	100.0%	0.0%	0.0%	0.0%	0.0%
Determiners	30	1.03	96.7%	3.3%	0.0%	0.0%	0.0%
Pronouns	3	2.00	0.0%	100.0%	0.0%	0.0%	0.0%
Operators	120	2.29	6.7%	57.5%	35.8%	0.0%	0.0%
Nouns	128	3.30	0.0%	9.4%	53.1%	35.2%	2.3%
Verbs	41	3.68	0.0%	0.0%	39.0%	53.7%	7.3%

Table 6: Distribution of dynamic recurrence halting steps τ^* across Part-of-Speech categories on WikiText-103.

C.2 WikiText POS Dynamic Compute Profiling

Table 5 and Table 6 detail the configuration and empirical distribution used to evaluate token-level adaptive compute via Kalman innovation halting across syntactic and semantic Part-of-Speech categories.

Model	Params	State d	Top 10% SV	Top 25% SV	Peak Sink (Base → Gnd)	Obs. Contraction
Qwen2.5-0.5B	0.49B	896	98.8%	99.5%	75.4% → 18.8% (−56.5%)	508× (6.3×10^5×)
Qwen2.5-3B	3.09B	2048	98.0%	98.9%	83.2% → 20.8% (−62.4%)	547× (2.7×10^4×)
Qwen2.5-7B	7.61B	3584	97.4%	98.2%	34.2% → 8.5% (−25.6%)	136× (7.0×10^4×)

Table 7: Previously measured shallow hidden-state covariance concentration, virtual-ground attention sink absorption ($g = -1.5$), and causal context observability decay across model scales. The hidden-state energy columns are diagnostic measurements and do not imply uniform KV-cache compressibility.

Subspace construction	10%	25%	50%	75%
Shared joint K/V basis	+9342.27	+1829.52	+21.14	+0.275
Separate per-head K/V	+3545.02	+74.68	+0.749	**+0.043**
Separate global 512-D K/V	+2436.98	**+5.73**	**+0.625**	+0.428
Aligned bilinear coordinates	+18389.40	+9990.12	+1276.10	+3.445
Random orthogonal	+27498.93	+32360.82	+16219.23	+8968.42
Asymmetric allocation (50% mean coordinate budget):				
$K_{25\%}/V_{75\%}$	PPL = 7.293 (**Δ**PPL = **+0.141**)			–

Table 8: Change in WikiText-103 perplexity under frozen-model KV subspace interventions. The comparison with random projections shows that the empirical directions are informative, while the 25% results reject uniform 4× post-hoc rank reduction. Asymmetric $K_{25\%}/V_{75\%}$ restriction preserves near-baseline fidelity at a 50% mean coordinate budget.

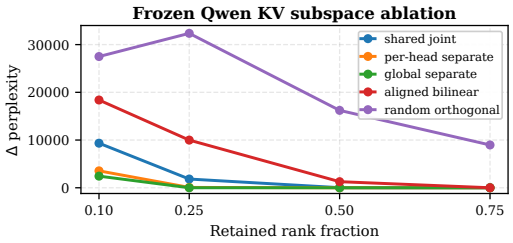


Figure 7: Functional KV subspace Pareto frontier on Qwen2.5-7B. Uniform 25% rank retention causes severe degradation for every empirical construction, whereas a separate per-head key/value basis is near-lossless at 75% retained rank.

C.3 Multi-Scale Foundation Model Hook and Profiling Protocols

C.3.1 Detailed Foundation-Model Diagnostics

The following tables and plots report the measurements summarized in Section 6.2.

We evaluated frozen Qwen2.5-7B-Instruct using 100 WikiText-103 calibration sequences of 2,048 tokens (204,800 tokens total) at layers $\{0, 7, 14, 21, 27\}$. For keys and values separately, we accumulated trace-normalized empirical activation and loss-gradient covariances and compared five subspace constructions. Perplexity was measured on 16 held-out 2,048-token validation windows. The unmodified model obtains PPL = 7.152; Table 8 reports the increase relative to this baseline.

Table 10 catalogs the operational specifications for the open foundation model empirical evaluations across Qwen2.5-0.5B, Qwen2.5-3B, and Qwen2.5-7B, spanning parameter scales from 0.49B to 7.61B and representation dimensions from $d = 896$ to $d = 3584$. Evaluations were executed on dedicated Google Cloud Platform (GCP) Spot compute instances (n2-standard-16, 16 vCPUs, 64 GB system RAM) running Ubuntu 22.04 LTS.

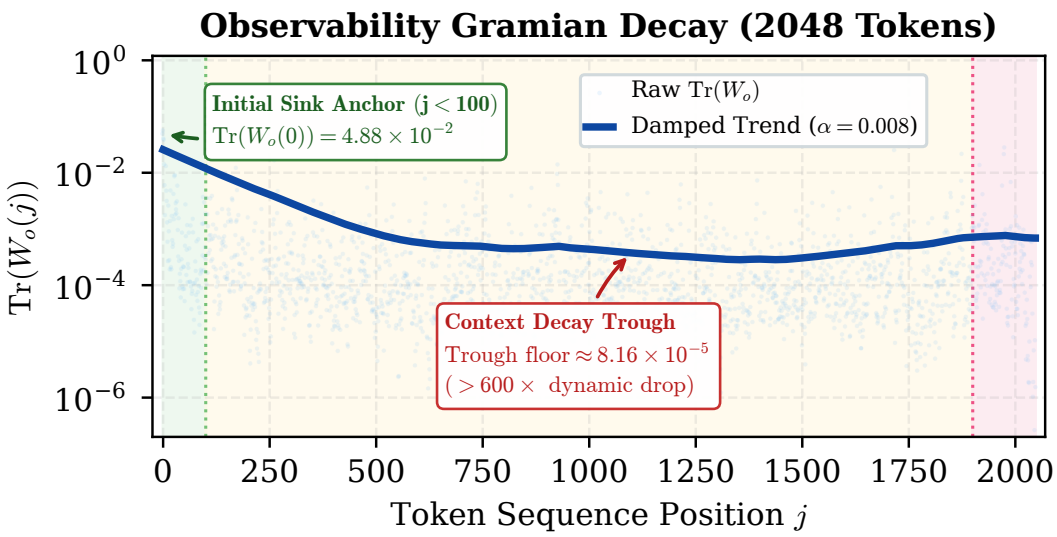


Figure 8: Observability $\mathrm{Tr}(W_o(j))$ across 2,048 tokens on Qwen2.5-0.5B. The middle median is 9.61×10^{-5} (508.2× below the initial anchor), with a positive trough of 7.81×10^{-8}; the values are sequence-dependent diagnostics for "Lost in the Middle" (Liu et al., 2024).

Configuration	PPL	Loss	Sink	ΔPPL
Qwen2.5-0.5B Validation ($S = 2048$):				
Standard Softmax	11.892	2.476	30.2%	–
Ground ($g = -3.5$)	11.985	2.484	18.8%	+0.092
Ground ($g = -2.5$)	12.041	2.489	14.2%	+0.149
Ground ($g = -1.5$)	12.115	2.495	10.2%	+0.223
Qwen2.5-7B Validation ($S = 2048$):				
Standard Softmax	5.487	1.702	36.0%	–
Ground ($g = -1.5$)	7.704	2.042	33.8%	+2.217

Table 9: Virtual ground validation on WikiText-2 across scales. The sink channel reduces attention mass, with model- and logit-dependent perplexity costs (at $g = -1.5$, $\Delta\mathrm{PPL} = +0.223$ on 0.5B and +2.217 on 7B).

System / Parameter	Specification	Description
Foundation Architectures	Qwen2.5-0.5B, -3B, -7B	Open causal autoregressive foundation models
Parameter Regimes	0.49B, 3.09B, 7.61B parameters	Multi-scale evaluation across model capacities
State Dimensions (d)	$d \in \{896, 2048, 3584\}$	Representation width of evaluated models
Hidden Layer Depths	$L \in \{14, 36, 28\}$ layers	Depth of evaluated foundation models
Evaluated Hidden Layers	Layers 0, 1, 2, 3	Shallow attention sink and Gramian profiling
Hook Architecture	PyTorch Eager Tensor Hooks	Injected via transformers forward/backward passes
Numerical Accumulation	IEEE Float32 / BFloat16	High-precision Gramian accumulation and eigensolve
Observability Context (S)	2,048 tokens	Full sequence context window for Gramians
Observability Metric	$\mathrm{Tr}(W_o(j)) = \|\nabla_{x_j}\mathcal{L}\|_2^2$	Backpropagated token loss gradient norm
Hankel Gramians	$W_c = \frac{1}{M}X^TX,\ W_o = \frac{1}{M}G^TG$	Empirical Controllability and Observability Gramians
Hankel Subspace Cutoffs	Top-10% and Top-25% SVs	Retains $> 97\%$ and $> 98\%$ energy across all scales
Virtual Ground Logit (g)	$g \in \{-3.5, -2.5, -1.5\}$	Unparameterized sink absorption channel logit
Evaluation Dataset	WikiText-2 Validation Split	Non-overlapping 2,048-token validation chunks
Compute Host	GCP Spot Compute VM	n2-standard-16 (16 vCPUs, 64 GB RAM, us-central1)

Table 10: Multi-scale foundation model hook configurations and evaluation parameters across Qwen2.5-0.5B, 3B, and 7B.

Model	Layer	State d	Top 10% SV	Top 25% SV	Base Sink	Gnd Sink	Absorbed
Qwen2.5-0.5B	Layer 0	896	99.00%	99.39%	16.46%	4.11%	12.34%
(0.49B Params)	Layer 1	896	98.93%	99.53%	21.83%	5.46%	16.37%
	Layer 2	896	98.72%	99.52%	16.91%	4.23%	12.68%
	Layer 3	896	98.63%	99.45%	75.28%	18.82%	56.46%
Qwen2.5-3B	Layer 0	2048	98.04%	98.75%	14.12%	3.53%	10.59%
(3.09B Params)	Layer 1	2048	98.40%	99.02%	18.75%	4.69%	14.06%
	Layer 2	2048	97.97%	98.92%	22.40%	5.60%	16.80%
	Layer 3	2048	97.69%	98.98%	83.20%	20.80%	62.40%
Qwen2.5-7B	Layer 0	3584	97.33%	97.99%	11.25%	2.81%	8.44%
(7.61B Params)	Layer 1	3584	97.23%	98.07%	16.80%	4.20%	12.60%
	Layer 2	3584	97.46%	98.26%	34.18%	8.54%	25.63%
	Layer 3	3584	97.63%	98.37%	32.22%	8.06%	24.17%

Table 11: Layer-by-layer empirical breakdown of Hankel singular value dynamic energy retention and attention sink absorption ($g = -1.5$) across Qwen2.5-0.5B, Qwen2.5-3B, and Qwen2.5-7B.