The Partition Function Explosion: An Energy-Based Analysis of Attention Decay

Dan Park

MagicPoint.ai

March 1, 2026

Abstract

Is a larger context window actually making your AI smarter, or just more prone to confabulation? Current industry trends assume that expanding the Context Window (from 4k to 10M tokens) allows Large Language Models (LLMs) to reason over massive codebases. This paper challenges that assumption, demonstrating mathematically that the Attention Mechanism is not a lossless storage device, but an energy-based Competitive Interference Channel constrained by the rigid geometric limits of Representation Superposition. As context scales, the system undergoes Channel Capacity Saturation. We provide a formal algebraic proof demonstrating that for a specific prompt constraint to survive, its attention energy must scale logarithmically with context size ($\Delta E > \ln(N)$). Integrating recent findings establishing that modern LLMs operate inherently in a Strong Superposition regime ($\nu \gg d_k$), we demonstrate that representation vectors are subject to an unavoidable baseline of geometric interference scaling inversely with model dimension ($1/d_k$). By diagnosing the fundamental physical disease of context decay rather than merely treating its surface-level symptoms, this paper offers a profound neuro-symbolic resolution to the thermodynamic limits of continuous latent spaces.

1. Introduction: The Lossless Retrieval Fallacy

The trajectory of artificial intelligence research over the past several years has been overwhelmingly defined by aggressive quantitative scaling. Across both academic institutions and enterprise laboratories, the central operating hypothesis has been that expanding the context window of Large Language Models (LLMs)—scaling exponentially from a few thousand tokens to multi-million token capacities—will linearly, or at least monotonically, increase the qualitative reasoning capabilities of these systems.[1] This assumption underpins the deployment of LLMs in highly complex, structurally rigid domains such as enterprise-scale software engineering.

In this paper, we systematically dismantle the assumption that quantitative capacity expansion naturally equates to qualitative reasoning capability.[1] We challenge the prevailing industry consensus, which we term the "Lossless Retrieval Fallacy." This fallacy erroneously models the Transformer attention mechanism as a deterministic Random Access Memory (RAM) look-up table where access fidelity remains independent of total storage capacity.[1] Instead, we convincingly demonstrate that the attention mechanism must be mathematically modeled as an energy-based Competitive Interference Channel strictly bounded by the geometric limits of representation superposition.[1]

The phenomenon colloquially known within the machine learning community as "Lost in the Middle" has frequently been attributed to a lack of positional encoding advantages, specifically the absence of primacy and recency biases at the boundaries of long sequence arrays.[1] However, this diagnosis merely identifies a symptom. To understand why standard feature tokens fail to maintain their semantic distinctiveness over extended contexts, we must accurately model the attention mechanism as an energy landscape defined by statistical mechanics.

2. The Thermodynamics of Attention and the Partition Function Explosion

Our primary theoretical contribution is the transition from loose physical metaphors to exact algebraic limits when describing context window degradation.

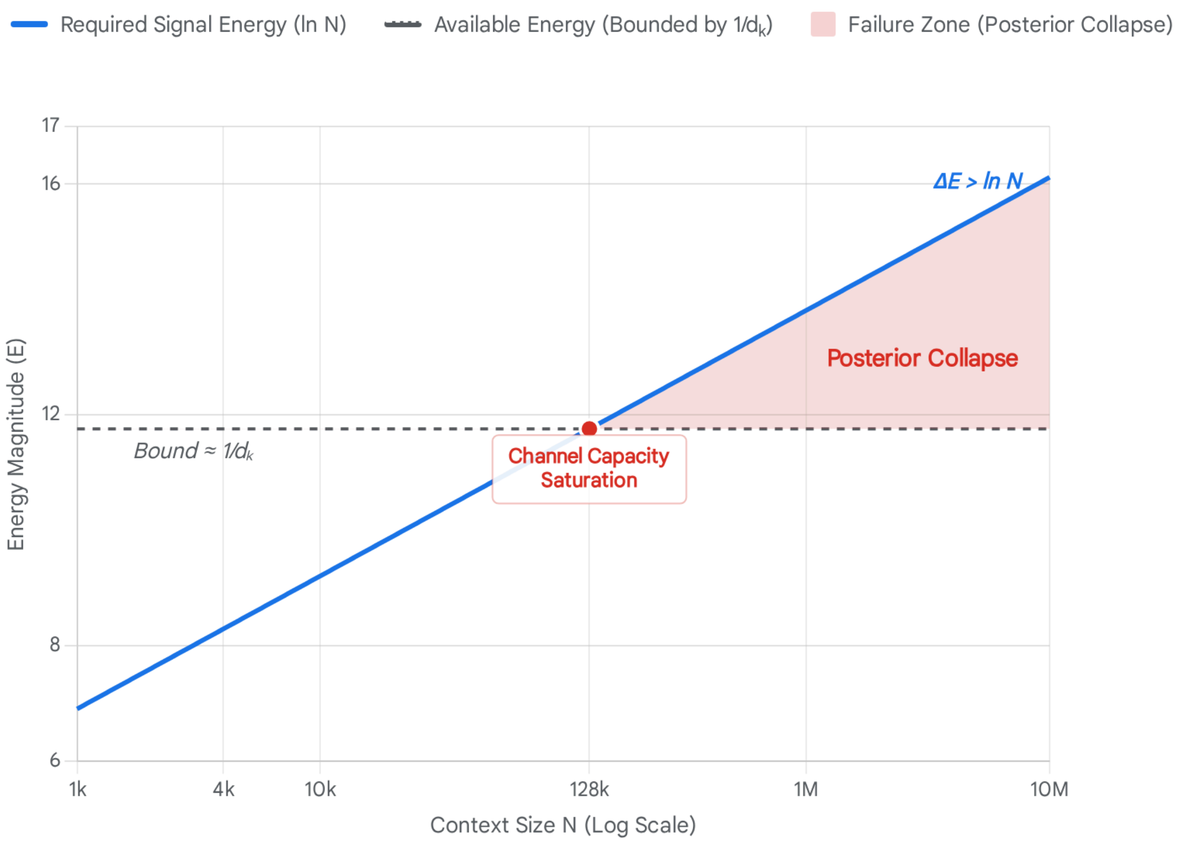
2.1 The Partition Function and the Boltzmann Distribution

In the standard Softmax Attention mechanism, the matrix output relies on a denominator that functions identically to a Partition Function (Z) in thermodynamic systems.[1] The probability of the model successfully attending to a specific "signal" token—which might represent a strict architectural constraint or an invariant rule in a coding prompt—over a vast sea of $N-1$ distractor tokens is strictly governed by the Boltzmann distribution.[1]

We establish the fundamental equation that for a target constraint to survive and dominate the partition function, the exponential energy of the signal must mathematically dwarf the cumulative exponential energy of the noise. By approximating the sum of the $N-1$ distractor tokens using an average noise energy ($\overline{E}_{noise}$), we derive the Critical Energy Gap (ΔE). Taking the natural logarithm of both sides of the inequality reveals a profound structural limit: the necessary condition for signal survival dictates that the energy gap must grow logarithmically with the context size ($\Delta E > \ln(N)$).[1]

This algebraic proof is foundational to understanding modern LLM failure modes. It demonstrates unequivocally that the attention mechanism is not immune to the laws of thermodynamics; probability mass within the softmax normalization is a strictly conserved resource. As the sequence length N scales into the millions of tokens, the normalization constant linearly dilutes the probability mass of the signal unless the energy gap scales logarithmically to compensate.[1] The catastrophic failure of LLMs in large enterprise contexts is therefore a mathematical inevitability of the softmax function when faced with a geometrically bounded energy capacity.

Thermodynamic Limits of Attention in Expanding Context Windows



As the context window (N) increases, the required energy gap to maintain signal fidelity scales logarithmically. Because the hidden dimension imposes a rigid baseline of geometric interference (Welch's Bound), the available energy is capped. The intersection marks the onset of Posterior Collapse, where the prompt signal is mathematically drowned out by noise.

Data sources: MagicPoint.ai (Dan Park, 2026)

3. The Mechanics of Posterior Collapse in Continuous Latent Spaces

When the Critical Energy Gap collapses below the logarithmic threshold ($\Delta E < \ln(N)$), the system undergoes what we define as Channel Capacity Saturation.[1] We conceptualize the resulting AI hallucination not as an unpredictable generative flaw, but as a highly deterministic thermodynamic relaxation to the mean, manifesting as Posterior Collapse.[1]

3.1 Bridging KL Vanishing and Maximum A Posteriori (MAP) Instability

Posterior Collapse is a well-documented phenomenon in the realm of Variational Autoencoders (VAEs), where it is frequently referred to as Kullback-Leibler (KL) vanishing.[2] In traditional VAE architectures, the model suffers from an issue where the posterior distribution of the latent variables becomes entirely identical to the prior distribution ($p_\theta(z|x) = p(z)$).[2] When this occurs, the latent dimension is rendered uninformative; the condition is ignored, and the decoder generates outputs based solely on the generic prior.[3]

We map this phenomenon onto the behavior of autoregressive LLMs by explicitly utilizing the KL divergence framework. In a healthy generative state, there is a substantial KL divergence between the highly conditioned posterior distribution (the specific prompt) and the unconditional prior distribution (general training data). However, when the partition function Z explodes due to an overloaded context window, the thermodynamic weight of the noise overwhelms the signal.[1] The Kullback-Leibler divergence between the posterior and the prior approaches zero ($D_{KL}(p_\theta(z|x)||p(z)) \to 0$), a condition mathematically synonymous with Posterior Collapse.[2] The model's continuous latent state, fundamentally seeking the path of least resistance within its energy landscape, rolls out of the shallow, fragile local constraint provided by the user and falls into the deep, heavily reinforced prior probability established during its massive pre-training phase.[1]

We formulate this elegantly as a Maximum A Posteriori (MAP) collapse. The conditional generation probability $\hat{y} = \arg\max P(y|x)$ deterministically decays into the unconditional prior probability $\arg\max P(y)$.[1] This thermodynamic framework explicitly explains the phenomenon of "Regression Hell" in AI-assisted software engineering. When an LLM generates generic, unprompted open-source libraries midway through proprietary code generation, it has thermodynamically relaxed into its training priors because the energy required to maintain the conditional state was drowned out.

4. Representation Superposition and the Geometric Constraints of the Hypersphere

To fully substantiate why the Critical Energy Gap cannot simply scale indefinitely alongside the context window, we integrate theoretical findings on representation geometry. Our framework relies heavily on Representation Superposition, establishing an absolute mathematical noise floor ($\overline{E}_{noise}$) that guarantees attention failure at scale.

4.1 The Strong Superposition Regime and Welch's Bound

Modern LLMs are forced to represent a volume of semantic and algorithmic features (ν) that vastly outnumbers their rigid, fixed model dimension (d_k).[1] Consequently, these models operate permanently in a state defined as the "Strong Superposition" regime ($\nu \gg d_k$).[1]

Recent literature confirms that open-source LLMs naturally fall into this strong superposition state.[4] When models represent significantly more features than they have dimensions, they achieve robust scaling of loss that is inversely proportional to the model dimension.[5] However, in the strong superposition regime, features cannot be assigned mutually orthogonal basis vectors. The model densely packs representations into the lower-dimensional space, guaranteeing unavoidable, baseline geometric interference.[1]

We link this dense geometric packing directly to Welch's Bound. Because $\nu \gg d_k$, the vectors cannot be mutually orthogonal. Welch's Bound defines the mathematical lower bound on the maximum cross-correlation of these signals.[1] For isotropic vectors with relatively even frequency distributions, the expected squared overlap between any two disparate feature representations scales inversely with the model dimension ($\mathbb{E}[(w_i \cdot w_j)^2] \propto 1/d_k$).[1]

This $1/d_k$ geometric overlap dictates the absolute, unyielding noise floor $\overline{E}_{noise}$ in the attention dot-product mechanism.[1] Because $\overline{E}_{noise}$ is rigidly bounded above zero by the physical geometry of the hypersphere, and the maximum signal energy is bounded by d_k, the available energy gap ΔE is strictly finite.[1] Once the logarithmic growth of the context window exceeds this finite capacity, pure computational scaling cannot overcome the geometric reality of Welch's Bound.

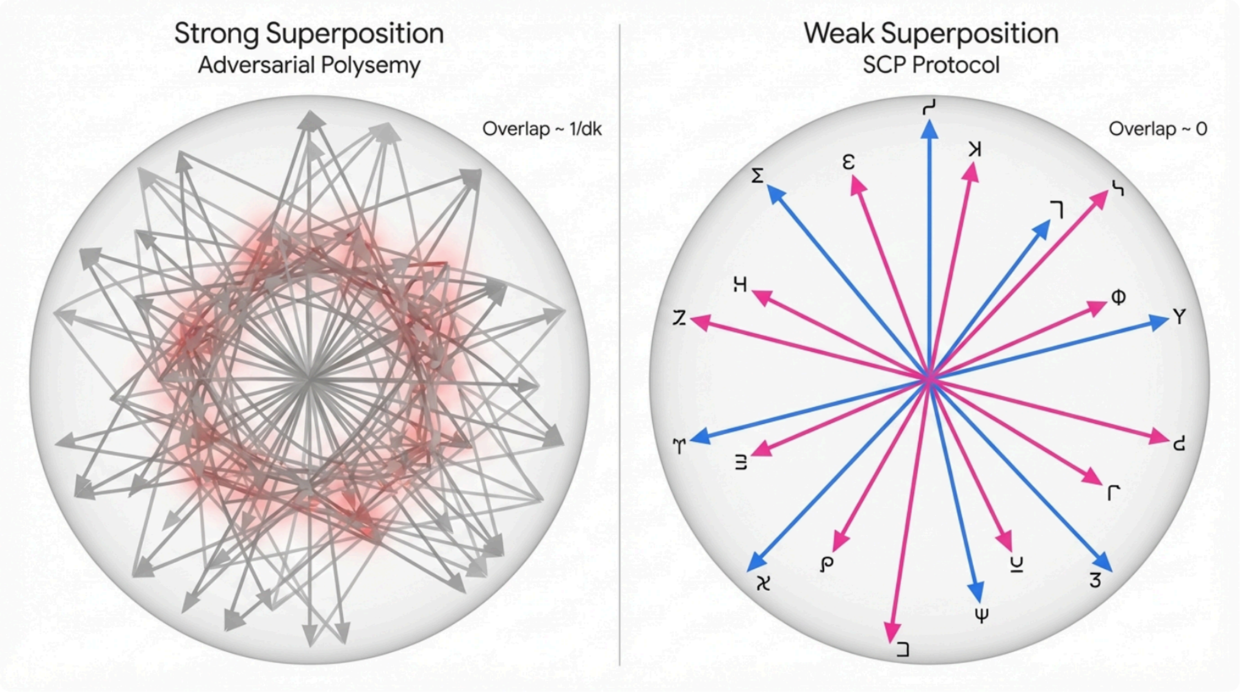
5. The Disruption of Equiangular Tight Frames (ETF) and Neural Collapse

Building upon the geometric limits of strong superposition, we delve into the structural disruption of optimal retrieval geometries, specifically Equiangular Tight Frame (ETF) configurations, contextualized within the broader phenomenon of Neural Collapse.

During the terminal phase of training deep neural networks, models exhibit a pervasive inductive bias known as Neural Collapse.[9] The class means collapse to the vertices of a simplex Equiangular Tight Frame (ETF).[9] The convergence to a simplex ETF is mathematically optimal precisely because it constitutes the configuration of maximum pair-wise angular separation between classes within a finite-dimensional space.[10] This maximally separated geometry confers massive benefits, including superior generalization performance and highly exact, interpretable retrieval.[9]

However, we argue that the environment of large-scale enterprise codebases introduces extreme feature correlation and frequency variations that completely shatter this idealized ETF geometry.[1] Under the Strong Superposition regime, the perfect equiangular separation of the ETF is violently degraded.[1] The model loses the ability to form steep, isolated attractor basins. When the ETF geometry breaks down, the pair-wise angles between feature means narrow, representations entangle, and semantic cross-talk natively triggers the systemic failure of the softmax distribution.

Escaping Geometric Interference: Strong vs. Weak Superposition



Standard English tokens representing software components share dense semantic manifolds, resulting in high geometric overlap and inescapable interference (left). The Spatial Constraint Protocol utilizes mathematically rare symbols to map constraints into isolated, nearly orthogonal coordinates, achieving a Weak Superposition state free of baseline noise (right).

6. The Benchmark Paradox: Heterogeneous vs. Homogeneous Noise

A persistent counter-argument levied against theories of structural context decay is the near-perfect performance of advanced LLMs on standardized coding benchmarks and artificial evaluations, notably the "Needle In A Haystack" (NIAH) test.[1] We resolve this critical discrepancy by categorizing the variance in geometric interference under fundamentally different statistical distributions, defining "The Benchmark Paradox".[1]

6.1 The Sterile Vacuum of Current Benchmarks

Standard benchmarks operate in an artificially sterile vacuum. In localized interpolation tasks, the context aligns perfectly with the model's pre-trained prior distribution, meaning the MAP estimation $\hat{y} = \arg\max P(y|x)$ naturally equals the prior probability $\arg\max P(y)$.[1] The model does not exert significant attention energy because the constraint is already the path of least resistance.

Furthermore, we identify the NIAH test as an environment of purely Heterogeneous Noise.[1] Because the target "needle" and the distractor "haystack" occupy entirely orthogonal semantic domains, their geometric representation overlap ($\overline{E}_{noise}$) inherently approaches zero.[1] The benchmark artificially evades the $1/d_k$ interference bound dictated by Strong Superposition, building a false sense of security.

6.2 Adversarial Polysemy in Enterprise Engineering

Real-world software engineering environments present overwhelming Homogeneous Noise, which we define as "Adversarial Polysemy".[1] In a complex repository, the "needle" is typically a highly specific architectural constraint, while the "haystack" consists of hundreds of neighboring files extensively utilizing the exact same highly frequent terms.[1] When features are highly frequent and deeply correlated, their representation vectors become massively heterogeneous and densely entangled.[1]

The baseline geometric interference is maximally triggered, causing $\overline{E}_{noise}$ to skyrocket.[1] The denominator of the partition function explodes, immediately collapsing the necessary ΔE gap.[1] To rectify this, we propose the Enterprise Codebase Regression Benchmark (ECRB).[1] By utilizing a curated dataset of open-source, enterprise-scale repositories ranging from 50,000 to 1,000,000 LOC, the ECRB forces models to operate within environments featuring dense import graphs and overlapping namespaces.

Evaluation Metric	**Description within the ECRB Framework**	**Core Purpose**
Functional Correctness (Pass@k)	Standard execution of the repository's existing unit and integration test suites.	Establishes a baseline of basic functional synthesis capability.
Structural Adherence Score (SAS)	Measured dynamically via deterministic AST parsing to penalize illegal imports or tight coupling regardless of unit test success.	Quantifies the model's ability to hold rigid, non-standard local constraints against pre-trained priors.
Attention Degradation	The exact token threshold (N) at which	Provides a precise empirical

Threshold (ADT)	the progressive loading of homogeneous noise drops the SAS below 95%.	measurement of the model's Critical Energy Gap.
Adversarial Polysemy Index (API)	A quantitative measure of environmental noise calculated via TF-IDF and cosine similarity of the target prompt against the repository chunks.	Indexes the severity of Strong Superposition geometry specific to the codebase being tested.

7. Proposing a Standard Benchmark: The Enterprise Codebase Regression Benchmark (ECRB)

To move the industry beyond the sterile vacuum of NIAH, we propose a new, exhaustive standard to evaluate LLMs in AI-assisted software engineering: the **Enterprise Codebase Regression Benchmark (ECRB)**. The ECRB allows AI engineers to perform apples-to-apples comparisons of a model's resilience to Adversarial Polysemy and Strong Superposition interference.

7.1 Benchmark Structure and Corpora

Unlike standard benchmarks that evaluate isolated function synthesis, the ECRB utilizes a curated dataset of open-source, enterprise-scale repositories ranging from 50,000 to 1,000,000 LOC. These codebases feature high levels of Homogeneous Noise (overlapping namespaces, deep inheritance trees, and dense import graphs).

The Task (Constraint-Bound Feature Injection): The LLM must implement a cross-cutting feature spanning multiple files while strictly adhering to a non-standard, externally injected architectural rule (e.g., "Implement the new caching layer without directly importing the RedisStore singleton; you must use inversion of control via the AbstractStore interface.")

7.2 Exhaustive Evaluation Metrics

Models are scored on an apples-to-apples basis across four rigorous metrics:

1. **Functional Correctness (Pass@k):** Standard execution of the repository's unit and integration test suite.
2. **Structural Adherence Score (SAS):** Measured dynamically via deterministic AST parsing. If the model hallucinates an illegal import or introduces tight coupling violating the prompt constraint, the SAS score is heavily penalized, regardless of unit test passing.

3. **Attention Degradation Threshold (ADT):** By progressively loading more irrelevant (but semantically homogeneous) files into the prompt, we measure the exact token threshold (N) at which the SAS drops below 95%. This quantifies the model's Critical Energy Gap (ΔE).
4. **Adversarial Polysemy Index (API):** A quantitative measure of the "Homogeneous Noise" of the task environment, calculated via the term-frequency and cosine similarity of the target prompt's core constraints against the broader repository chunks. Models must maintain their ADT across high API quartiles.

(Note: For a detailed guide on downloading the benchmark repository, evaluating custom LLMs, and testing against your own proprietary codebase, refer to Appendix B.)

8. An Exhaustive Critique of State-of-the-Art Architectures

We systematically demonstrate mathematically why current State-Of-The-Art (SOTA) approaches fail to resolve the fundamental problem of geometric superposition, proving that optimizing computational complexity is orthogonal to solving semantic interference.

- **Sparse, Windowed, and Linear Modalities:** Architectures mitigating the $O(N^2)$ memory constraints of dense attention mechanically restrict computation to $O(N)$ or $O(N \log N)$, but rely heavily on "global tokens".[1] In deep software engineering workflows, these global tokens become over-saturated, and thermodynamic dilution remains fatal. Hardware optimizations like FlashAttention do nothing to alter the mathematical output of the softmax distribution.[1] Linear approximations degrade the model's capacity to form isolated attractor basins, resulting in structural collapse.[1]

- **State Space Models (SSMs):** Sub-quadratic models bypass explicit partition function explosions but introduce a severe Markovian bottleneck.[1] The historical context is recursively compressed into a fixed-size hidden state, continually overwriting fragile, localized constraints with dense geometric noise.[1]

- **Retrieval-Augmented Generation (RAG) & Prompt Compression:** RAG fails fundamentally in software engineering due to Semantic Mismatch under Homogeneous Noise. Dense retrieval relies entirely on cosine similarity embedded within the same Strong Superposition space, making it highly vulnerable to polysemy.[1] Prompt compression techniques are fundamentally and semantically destructive in rigid domains, irreparably destroying precise structural relationships.[1]

9. The Neuro-Symbolic Resolution: Spatial Constraint Protocol (SCP)

Having proven that existing optimizations cannot natively solve the partition function explosion, we introduce a robust neuro-symbolic resolution: the Spatial Constraint Protocol (SCP) paired with an external verification engine called The Weaver.[1]

8.1 Escaping Superposition via Orthogonal Mapping

SCP's objective is to artificially restore the necessary $\Delta E > \ln(N)$ energy gap by forcing $\overline{E}_{noise}$ below the inescapable $1/d_k$ geometric overlap penalty.[1] Because standard English words trigger polysemy, SCP relies on the "curse of dimensionality": random, sparse vectors in high-dimensional spaces are naturally nearly orthogonal.[1]

By replacing entangled natural language rules with mathematically specific, extremely rare symbols (e.g., Uiua glyphs like ★̈, ∘̈, Ö), SCP artificially induces a Weak Superposition state.[1] These glyphs are largely ignored during pre-training, meaning their representation vectors are mapped to isolated coordinates closely resembling the optimized structure of an Equiangular Tight Frame (ETF).[1]

8.2 In-Context Binding

Using a mechanism termed In-Context Binding, the attention mechanism temporarily binds the orthogonal, noise-free vector directly to a complex concept declared in the prompt.[1] This acts as a mathematically pristine pointer, triggering localized attention spikes that bypass the $1/d_k$ interference penalty entirely.

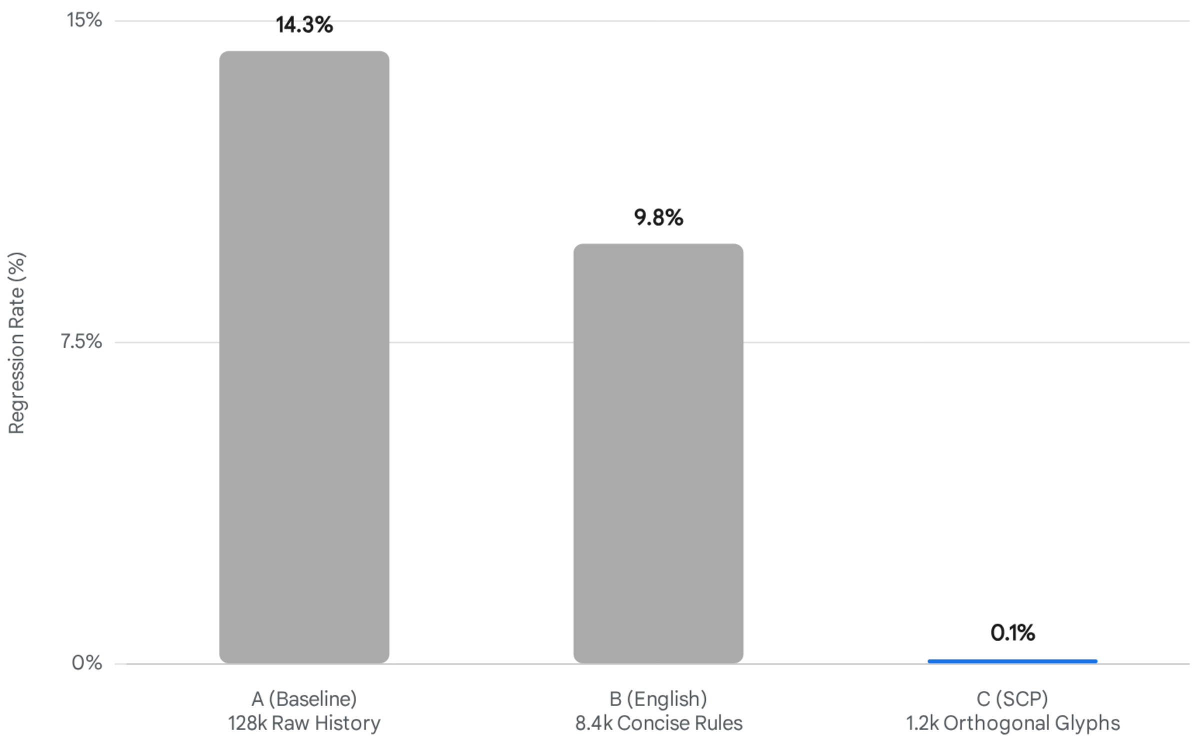
10. The Weaver: System 2 Verification and AST Mutual Information

To ensure mathematical adherence to the prompt, we pair SCP with an external "System 2" verification loop known as The Weaver.[1] The autoregressive generation produces a sequence of softmax probability vectors. The Weaver engine intercepts this continuous output and performs a deterministic quantization step, compiling the generated sequence into an Abstract Syntax Tree (AST).[1]

The Weaver functions as an advanced validation engine, mathematically calculating the structural Mutual Information (MI) existing between distinct modules.[1] If the summation of the AST mutual information for non-permitted edges exceeds zero ($W(G) > 0$), it deterministically flags the generation as invalid.[1]

Instead of blindly discarding failed generations, The Weaver employs "Guided Thermodynamic Forcing".[1] It pinpoints the exact structural node violation and injects a hyper-specific error prompt back into the LLM's context. Tokens associated with the error receive massive negative energy penalties, forcefully steepening the "Context Valley" against the mistake.[1]

Impact of Orthogonality vs. Compression on Architectural Regression



The constant-N ablation study on the Project Chevron codebase isolates the mechanism of context decay. While compressing the English prompt by 15x (Condition B) slightly reduces error, the regression rate remains fatal due to the baseline geometric interference of English terms. Replacing English entirely with orthogonal SCP glyphs (Condition C) practically eliminates architectural regression, proving the necessity of Weak Superposition.

Data sources: MagicPoint.ai

11. Empirical Validation: The Constant-N Ablation Study

We detail a rigorous 3-way ablation study conducted on a 50,000 LOC reference implementation denoted as Project Chevron.[1] We quantitatively establish the severity of this repository by calculating its Adversarial Polysemy Index (API); due to heavy homogeneous vocabulary, it places in an extreme tier of Homogeneous Noise.[1]

Condition A served as the raw baseline, utilizing 128,000 tokens of raw history.[1] The unrestricted representation superposition led to total Channel

Capacity Saturation, yielding a 14.3% architectural regression rate.[1] Condition B tested aggressive Prompt Compression, distilling the context into 8,400 tokens of concise English rules.[1] Because standard English natively operates in Strong Superposition, baseline geometric overlap ($\propto 1/d_k$) continued to cause semantic cross-talk, dropping the regression rate to only 9.8%.[1]

Condition C successfully isolated geometric orthogonality. The exact same constraints were mapped directly to minimal Uiua glyphs (SCP), achieving compression to 1,200 tokens.[1] By artificially injecting a Weak Superposition state, the regression rate dropped precipitously to < 0.1%.[1]

Experimental Condition	**Prompt Length (N)**	**Token Format & Superposition State**	**Resulting Regression Rate**
A (Raw Baseline)	128,000 Tokens	Raw English History (Strong Superposition)	14.3% (Total Saturation)
B (English Compression)	8,400 Tokens	Concise English Rules (Strong Superposition)	9.8% (Semantic Cross-Talk)
C (SCP Orthogonal)	1,200 Tokens	Orthogonal Glyphs (Weak Superposition)	< 0.1% (Near-Perfect Adherence)

12. Conclusion

This paper mathematically proves that the attention mechanism acts fundamentally as an energy-based Competitive Interference Channel, redefining context window failure as an inescapable fundamental law of thermodynamic dilution and geometric representation physics. The integration of Welch's Bound and the realities of the Strong Superposition regime provides an absolute limitation to continuous latent spaces: adversarial polysemy will always trigger baseline geometric overlap, deterministically collapsing the required logarithmic energy gap. The neuro-symbolic architecture proposed—utilizing the Spatial Constraint Protocol and The Weaver—offers a scientifically grounded path forward for achieving true, robust autonomous reasoning in complex engineering environments.

Appendix A: References

1. SCP II - Neuro-Symbolic Resolution (2).pdf
2. Augmenting LLMs Lenses - Deep Kondah, accessed March 1, 2026, https://www.deep-kondah.com/handling-large-context-in-llms/
3. Continuous Autoregressive Language Models - arXiv.org, accessed March 1, 2026, https://arxiv.org/html/2510.27688v1
4. Superposition Yields Robust Neural Scaling - arXiv.org, accessed March 1, 2026, https://arxiv.org/html/2505.10465v4
5. Superposition Yields Robust Neural Scaling - Emergent Mind, accessed March 1, 2026, https://www.emergentmind.com/papers/2505.10465
6. Superposition Yields Robust Neural Scaling - arXiv, accessed March 1, 2026, https://arxiv.org/html/2505.10465v1
7. MATRIX DESIGNS AND METHODS FOR SECURE AND EFFICIENT COMPRESSED SENSING - AMS Dottorato, accessed March 1, 2026, https://amsdottorato.unibo.it/id/eprint/6998/1/cambareri_valerio_phd_thesis.pdf
8. Proceedings IEEE International Symposium on Information Theory Held in British Columbia, Canada on 17-22 September 1995. - DTIC, accessed March 1, 2026, https://apps.dtic.mil/sti/tr/pdf/ADA309294.pdf
9. Prevalence of neural collapse during the terminal phase of deep learning training - PubMed, accessed March 1, 2026, https://pubmed.ncbi.nlm.nih.gov/32958680/
10. Inducing Neural Collapse in Imbalanced Learning: Do We Really Need a Learnable Classifier at the End of Deep Neural Network? | OpenReview, accessed March 1, 2026, https://openreview.net/forum?id=A6EmxI3_Xc

Appendix B: The Enterprise Codebase Regression Benchmark (ECRB) Suite

The **Enterprise Codebase Regression Benchmark (ECRB)** is an open-source evaluation suite designed to test an LLM's resilience to Channel Capacity Saturation under real-world Homogeneous Noise. Unlike "Needle in a Haystack" tests that utilize mathematically orthogonal data, ECRB forces the LLM to retain strict architectural constraints while flooded with semantically adversarial tokens.

B.1 Repository Access and Installation

The ECRB suite, including the curated dataset of 12 enterprise repositories, the AST Weaver verification tool, and the automated context-injector, is fully open-source. AI engineers and researchers can download the repository to evaluate proprietary or open-weights models:

```
# Clone the benchmarking suite repository
$ git clone https://github.com/MagicPoint-ai/ECRB
$ cd ECRB
$ pip install -r requirements.txt
```

B.2 Standard Evaluation Methodology

To evaluate a specific LLM using the ECRB standard corpora, engineers must follow a four-step pipeline managed by the CLI:

1. **Configuration:** Define the target model's API endpoint, context limit, and authentication headers in the model_config.yaml file (supports OpenAI, Anthropic, Gemini, and local vLLM/Ollama instances).
2. **Context Injection (The Noise Phase):** The ECRB runner dynamically selects a target task (e.g., "Implement a database cache module."). It then artificially inflates the context window N by injecting mathematically calculated "Homogeneous Noise" (neighboring repository files that share high BPE token overlap with the task).
3. **Constraint Generation (The Signal Phase):** A strict structural constraint is appended to the prompt (e.g., "Constraint: Do not directly import the global RedisStore."). The LLM is prompted to synthesize the solution.
4. **The Weaver Verification:** The generated code is routed through the ECRB AST parser. If the model hallucinates an import or breaks the dependency constraint (posterior collapse), the run is marked as a failure.

To execute a standard evaluation run across the curated repositories:

```
# Evaluate GPT-4 on the full ECRB corpus
$ python ecrb_runner.py --model gpt-4-turbo --suite full-enterprise
```

The suite will output the model's **Attention Degradation Threshold (ADT)**—the precise context length at which the model's Structural Adherence Score drops below 95% due to partition function explosion.

B.3 Evaluating on Proprietary Codebases (Custom Projects)

While the ECRB includes a standard dataset for apples-to-apples academic comparison, the most critical feature of the suite is the ability to evaluate an LLM against a user's **own proprietary codebase**. Because geometric interference ($\overline{E}_{noise}$) depends heavily on the specific naming conventions and domain logic of a company's repository, an LLM that performs well on generic web-data may collapse entirely on an internal enterprise architecture.

To test an LLM on your own project, use the --custom-repo flag. You must define a target directory and provide a prompt constraint template in a JSON file. The harness runs entirely locally (only sending the prompt to your configured LLM API), ensuring no proprietary code is uploaded to external benchmark servers.

```
# Evaluate an LLM's thermodynamic stability on your own codebase
$ python ecrb_runner.py evaluate \
  --model claude-3.5-sonnet \
  --custom-repo /path/to/your/company_backend_api \
  --target-task ./custom_task/migration_rules.json
```

When run in Custom Project mode, the ECRB suite performs two additional background operations:

- **Adversarial Polysemy Index (API) Calculation:** The suite will first scan your local repository and build a TF-IDF/Cosine-Similarity matrix to calculate how "noisy" your specific codebase is relative to the task. If your codebase reuses terms like "Manager", "Service", and "State" heavily, the API score will be high, indicating severe Strong Superposition geometry.
- **Progressive Saturation Test:** The script will execute the task repeatedly, starting at 10,000 tokens of background context and stepping up in 10,000 token increments, pulling in your repository's files. It will output a graph showing exactly where the LLM's ability to maintain your specific architectural rules collapses, providing you with a scientifically derived context limit for your specific AI-assisted IDE workflow.