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

An Analysis of Representation Superposition, Geometric Interference, and Guided Thermodynamic Forcing

Dan Park
MagicPoint.ai

February 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 by Liu et al. (2025) which establish 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 ($1/d_k$). When the $\Delta E > \ln N$ boundary is breached, this cumulative geometric overlap mathematically drowns out the signal.

This forces the model into *Posterior Collapse*—abandoning the local constraints of the prompt to relax into lowest-energy pre-trained priors.

Following an exhaustive review of current State-of-the-Art (SOTA) transformer variants, we demonstrate that existing architectures optimize for computational complexity rather than this underlying geometric and thermodynamic limit. We explain the "Benchmark Paradox"—why LLMs succeed at 1-million-token "Needle In A Haystack" tests but regress catastrophically in real-world AI-assisted software engineering workflows due to *Adversarial Polysemy* disrupting Equiangular Tight Frame (ETF) configurations. To establish a rigorous ground truth, we propose a new open standard: the *Enterprise Codebase Regression Benchmark (ECRB)*.

Finally, we resolve the algorithmic limit by introducing the Spatial Constraint Protocol (SCP) paired with *The Weaver*. SCP escapes the Strong Superposition trap by utilizing mathematically rare, orthogonal symbols (Uiua glyphs) to simulate a Weak Superposition ETF limit, artificially inducing low-entropy, zero-interference retrieval. Through a controlled ablation study on a 50,000 LOC codebase, we isolate orthogonality from prompt compression, demonstrating a reduction in architectural regression from 14.3% to <0.1%.

1 Introduction: The Lossless Retrieval Fallacy

The trajectory of artificial intelligence research (2023–2026) has been defined by the aggressive expansion of the Context Window (N). From 4,096 tokens to 10 million, the industry operates under the tacit assumption termed the "Billion Token Fallacy"—that quantitative capacity expansion equates to qualitative reasoning capability [13]. This relies on the *Lossless Retrieval Fallacy*: the assumption that the attention mechanism functions as a deterministic RAM look-up table where access fidelity is independent of total capacity [8].

Critics frequently point to empirical observations like the *"Lost in the Middle"* phenomenon [9] as a simpler alternative explanation for context failure. However, "Lost in the Middle" merely describes the *symptom* of context decay; it fails to diagnose the underlying physical *disease*. Models lose data in the middle precisely because those tokens lack the positional encoding advantages (primacy and recency biases) found at the sequence boundaries. Lacking this artificial mathematical boost to their energy, their retrieval relies entirely on their semantic distinctiveness. To fundamentally understand why standard tokens fail to maintain this distinctiveness over long contexts, we must model Attention as an energy landscape bounded by the geometry of representation superposition.

2 Theoretical Framework: The Exact Physics of Attention Decay

To understand why "more context" leads to confabulation (e.g., "Regression Hell" in AI-assisted software engineering), we must transition from physical metaphors to direct algebraic limits and geometric representation theory.

2.1 Formal Proof of the Critical Energy Gap (ΔE)

In a standard Softmax Attention mechanism, the matrix output is computed as:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

Where the denominator of the softmax operation acts as the Partition Function (Z) over the context length N:

$$Z = \sum_{j=1}^{N} e^{\text{score}(q,k_j)} \tag{2}$$

The probability (P_{sig}) of attending to a specific "signal" token (the target constraint) over $N - 1$ distractor tokens is governed by the Boltzmann distribution:

$$P_{sig} = \frac{e^{E_{sig}}}{e^{E_{sig}} + \sum_{j \neq sig}^{N} e^{E_{noise,j}}} \tag{3}$$

(Where energy $E = \frac{QK^T}{\sqrt{d_k}}$)

Let us approximate the sum of the $N-1$ distractor tokens using an average noise energy $\bar{E}_{noise}$, such that the denominator becomes $e^{E_{sig}} + Ne^{\bar{E}_{noise}}$.

For the model to successfully retrieve the target constraint without hallucinating during AI-assisted software engineering tasks, the signal must mathematically dominate the partition function ($P_{sig} \approx 1$). Thus, we require:

$$e^{E_{sig}} \gg Ne^{\bar{E}_{noise}} \tag{4}$$

Taking the natural logarithm of both sides yields the necessary condition for survival:

$$E_{sig} > \ln(N) + \bar{E}_{noise} \tag{5}$$

This reveals the **Critical Energy Gap**:

$$\Delta E = E_{sig} - \bar{E}_{noise} > \ln(N) \tag{6}$$

2.2 The Superposition Bottleneck: Geometric Interference

Equation 6 proves that as N scales into the millions, the required energy gap ΔE must grow logarithmically. However, a Transformer's hidden dimension (d_k) is rigidly bounded. To understand why LLMs fail to maintain this gap, we must integrate recent findings on *Representation Superposition* [10].

Modern LLMs are forced to represent vastly more linguistic and algorithmic features (ν) than their model dimension (d_k), placing them permanently in the **Strong Superposition** regime ($\nu \gg d_k$). In this regime, features cannot be assigned orthogonal basis vectors. Instead, the model packs representations densely, resulting in unavoidable geometric interference. Liu et al. (2025) prove that the minimum maximum-overlap between ν unit vectors in d_k dimensions is rigidly constrained by Welch's bound [18]:

$$\max_{i \neq j} |w_i \cdot w_j| \geq \sqrt{\frac{\nu - d_k}{d_k(\nu - 1)}} \approx \sqrt{\frac{1}{d_k}} \tag{7}$$

For isotropic vectors with relatively even frequency distributions, the expected squared overlap between any two disparate feature representations scales inversely with the dimension:

$$\mathbb{E}[(w_i \cdot w_j)^2] \propto \frac{1}{d_k} \tag{8}$$

This $1/d_k$ geometric overlap dictates the absolute noise floor $\bar{E}_{noise}$ in the attention dot-product. Because $\bar{E}_{noise}$ is rigidly bounded above zero by the physics of Strong Superposition, and E_{sig} is bounded by d_k, the energy gap ΔE is strictly finite. Once $\ln(N)$ exceeds this finite capacity, the signal is mathematically guaranteed to be drowned out by the partition function explosion.

2.3 Confabulation as Posterior Collapse

"Regression Hell" in AI-assisted software engineering is a manifestation of Mode Collapse driven by this thermodynamic limit. When Z explodes and ΔE collapses below $\ln N$, the local distribution defined by the prompt becomes too high-entropy. The model's latent state, seeking the path of least resistance, rolls out of the shallow local constraint and falls into the deep prior probability established during pre-training. Hallucination is not a creative act; it is a deterministic thermodynamic relaxation to the mean (Maximum A Posteriori collapse) [17]:

$$\hat{y} = \arg\max_y P(y|x) \to \arg\max_y P(y) \tag{9}$$

3 Priors and Current Works: SOTA Architectures for Context Scaling

The academic and industrial communities have recognized the computational overhead and qualitative degradation associated with long contexts. However, an exhaustive investigation reveals that current State-Of-The-Art (SOTA) methodologies primarily address computational complexity ($O(N^2)$ memory constraints) or positional extrapolation, rather than resolving the core geometric interference and thermodynamic dilution of the signal under Strong Superposition.

3.1 Sparse, Windowed, and Structured Attention

Architectures such as **Longformer** [1] and **BigBird** [20] replace dense global attention with sparse, sliding-window, or random attention patterns, reducing computational complexity to $O(N)$ or $O(N \log N)$. While these mechanisms restrict the size of the attention matrix mechanically, they rely on designated "global tokens" for cross-document reasoning. When evaluating deeply nested architectural constraints in AI-assisted software engineering, these global tokens become over-saturated. The partition function (Z) still dilutes the probability mass across the globally attended indices. Hardware-aware optimizations like **FlashAttention** solve SRAM IO bottlenecks but do absolutely nothing to alter the mathematical output of Eq. 1; the thermodynamic dilution remains identically fatal.

3.2 Linear Attention and Kernel Methods

Approximate methods like **Linformer** and **Performer** project the $N \times d_k$ sequence into lower-dimensional spaces. While Linear Attention limits the growth of the partition function mechanically, it inherently sacrifices the exact, sharp associative recall capacity necessary for rigid software engineering constraints. By replacing the softmax with linear kernels, the ability to form steep, isolated attractor basins out of the Strong Superposition interference is severely degraded.

3.3 State Space Models (SSMs) and Linear RNNs

Advances in sub-quadratic sequence modeling—most notably **Mamba** (Selective SSMs) [4], **RetNet** [16], and **RWKV** [11]—replace softmax attention entirely with continuous-time differential equations, linear time-invariant systems, or recurrent formulations.

Mathematically, these models bypass the explicit partition function explosion because they do not compute a normalized sum over all past tokens. However, they introduce a severe *Markovian bottleneck*: the historical context must be compressed into a fixed-size hidden state. In environments with severe representation superposition, this fixed-capacity state struggles with exact discrete retrieval. The recursive updates continuously overwrite fragile local constraints with dense geometric noise, leading to catastrophic forgetting of architectural rules.

3.4 Positional Embedding Extrapolations

Methods modifying the positional encoding, such as **Rotary Position Embeddings (RoPE)** [15], dynamic scaling methods like **YaRN** [12], and **ALiBi**, focus on out-of-distribution sequence length extrapolation. They allow a model trained on 4k tokens to computationally process 128k tokens without catastrophic attention distribution failure. However, they address *how* the model indexes long context (positional syntax), not *what* it pays attention to. They do nothing to mitigate the semantic geometric overlap ($\bar{E}_{noise}$) dictated by Eq. 8.

3.5 Dense Retrieval and Retrieval-Augmented Generation (RAG)

RAG frameworks [7] and Dense Passage Retrieval bypass the context limit by externalizing memory into vector databases and truncating the prompt to the top-k nearest neighbors. In AI-assisted software engineering, RAG fails fundamentally due to *Semantic Mismatch* under Homogeneous Noise. Because dense retrieval relies on cosine similarity in a Strong Superposition space, hundreds of unrelated files sharing the polysemic vocabulary of "authentication" or "database" will crowd out the strict, global architectural invariant, leading to representation entanglement and top-k retrieval collapse.

3.6 KV Cache Eviction and Prompt Compression

Techniques like **StreamingLLM** [19] utilize "Attention Sinks" combined with rolling eviction to prevent KV-cache overflow. Similarly, **LLMLingua** [5] employs prompt compression to drop high-entropy tokens, while **AutoCompressors** [3] learn condensed continuous representations via gradient descent. While these approaches stabilize continuous generation in conversational AI, they are semantically destructive. In AI-assisted software engineering, dropping syntactical tokens or compressing text within a polysemic manifold destroys the precise, discrete structural relationships required to maintain architectural integrity.

4 The Benchmark Paradox: Heterogeneous vs. Homogeneous Noise

A frequent critique of context degradation theories is the phenomenal performance of LLMs on standardized coding benchmarks (e.g., SWE-bench, HumanEval) and artificial long-context evaluations like "Needle In A Haystack" (NIAH). If models can perfectly find a hidden password in 1 million tokens, why do they catastrophically confabulate when deployed in an AI-assisted IDE on a 50,000 LOC proprietary codebase? This discrepancy is perfectly explained by examining geometric interference under different distribution types.

4.1 The Sterile Vacuum: Why Benchmarks Succeed

Standard coding benchmarks evaluate models in artificially sterile environments that guarantee success through two physical mechanisms:

- **Aligned Priors (Local Interpolation):** Benchmarks often test generic algorithmic patterns. In these cases, the Context Valley and the Prior Canyon are aligned. The model relies on its pre-trained weights, where $\hat{y} = \arg\max P(y|x)$ naturally equals $\arg\max P(y)$.
- **Heterogeneous Noise (NIAH):** In a standard NIAH test, the "needle" (a secret password) is semantically completely distinct from the "haystack" (a massive essay). Because the target vectors and distractor vectors occupy completely orthogonal semantic domains, their representation overlap ($\bar{E}_{noise}$) approaches zero, successfully evading the $1/d_k$ geometric interference bound of strong superposition. The benchmark artificially manufactures a steep energy gap.

4.2 Adversarial Polysemy: Why Large Projects Fail

Real-world codebases represent the exact opposite of a NIAH benchmark.

In a 50k LOC enterprise project, the "needle" is a highly specific architectural constraint (e.g., "Do not directly access the `db_config` in the `Auth` module"). The "haystack" consists of hundreds of other files that *also* use the exact same terms: `db_config`, `Auth`, `user_state`, and `module`.

This creates **Homogeneous Noise** or **Adversarial Polysemy**. As Liu et al. (2025) note, when features are highly frequent and correlated, representation vectors become massively heterogeneous [10]. The idealized Equiangular Tight Frame (ETF) geometry breaks down. Because standard BPE token embeddings for these words overlap heavily, the $\sim 1/d_k$ baseline interference is maximally triggered and exceeded. The average noise energy $\bar{E}_{noise}$ skyrockets. The denominator Z (Eq. 2) explodes, collapsing the ΔE gap, and forcing the LLM to hallucinate general internet priors.

5 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.

5.1 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."*)

5.2 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**.)*

6 The Resolution: Spatial Constraint Protocol (SCP)

To restore $\Delta E > \ln(N)$ (Eq. 6) in the presence of homogeneous codebases, we must engineer $\bar{E}_{noise}$ to break out of the Strong Superposition $1/d_k$ geometric overlap penalty identified by Liu et al. (2025).

6.1 Escaping Superposition via ETF Optimization

Standard English words ("sort", "state", "module") suffer from massive *polysemy*. Because these tokens are highly frequent, they are mathematically forced to represent multiple varying features, permanently trapping them in the Strong Superposition regime.

However, the "curse of dimensionality" dictates that random, sparse vectors in high-dimensional spaces are naturally nearly orthogonal. We formalize this constraint to artificially induce the low noise of a sterile benchmark:

$$\mathbb{E}[\text{sim}(e_{SCP}, e_{distractor})] \approx 0 \tag{10}$$

The Spatial Constraint Protocol (SCP) replaces entangled natural language architectural rules with mathematically specific, extremely rare symbols (Uiua glyphs, such as ⍜ or ⊸). Because these glyphs are largely ignored or utilized sparsely during pre-training, they are not subjected to the intense semantic packing of the Strong Superposition manifold. They exist in a *Weak Superposition* limit, mapped to isolated, sparse, ETF-like coordinates on the hypersphere. By shifting the representation out of the densely packed English manifold, SCP drops the expected geometric overlap from $1/d_k$ to near zero, freeing the prompt constraint from structural linguistic interference.

6.2 In-Context Binding

SCP does not require the model to "know" what ⍜ means zero-shot. Instead, it relies on **In-Context Binding**. By defining an SCP constraint in the prompt (*"Let ⍜ = Strict Module Isolation"*), the attention mechanism temporarily binds the orthogonal vector to the required concept. Because ⍜ carries zero historical adversarial polysemy, it acts as a lossless pointer, triggering localized attention spikes (massive E_{sig}) that bypass the $1/d_k$ interference penalty entirely.

7 The Weaver: Guided Thermodynamic Forcing

Standard intuitive Transformers (System 1) cannot natively compute exact, discrete Mutual Information (MI) during a continuous forward pass. SCP addresses this via an external System 2 verification loop known as *The Weaver*.

7.1 Maxwell's Demon and AST Mutual Information

The Weaver extracts the Abstract Syntax Tree (AST) from generated code and calculates structural Mutual Information between modules:

$$W(G) = \sum_{(i,j) \notin E} MI_{AST}(m_i, m_j) \tag{11}$$

If $W(G) > 0$ (e.g., an illegal global import is detected), the generation is mathematically invalid.

7.2 Guided Thermodynamic Forcing

Pure rejection sampling is computationally unviable. The Weaver solves this via **Guided Thermodynamic Forcing**. We *never* blindly discard a generation. The deterministic AST parser pinpoints the exact structural violation and dynamically injects a hyper-specific error back into the LLM's prompt.

By appending text like `[SYSTEM 2 REJECTION]: Line 42 contains an illegal module coupling...` , the attention matrix is physically reshaped. Tokens associated with the error receive massive negative energy penalties, artificially steepening the Context Valley for the second attempt. Because SCP already provides steep orthogonal attractor basins by escaping Strong Superposition geometry, the baseline zero-shot acceptance rate is high (>94%), and Guided Forcing guarantees compliance in 1 to 2 targeted resamples, ensuring economic viability.

8 Empirical Validation: The Constant-N Ablation Study

To isolate geometric orthogonality from prompt compression length, we conducted a strict 3-way ablation study on the `Project Chevron` reference implementation (a ~50,000 LOC codebase).

- **Condition A (Raw Baseline):** 128,000 tokens of raw natural language history and context.
- **Condition B (English Compression):** 8,400 tokens. The 128k context was summarized into the most concise standard English architectural rules possible. This represents a ~15x compression. The prompt length of this English compression is **about 7 times** that of the SCP glyphs.

- **Condition C (SCP Orthogonal Compression):** 1,200 tokens. The exact same constraints were mapped to minimal Uiua glyphs (e.g., `⍜(A, B)`) using In-Context Binding, achieving >100x compression.

8.1 Results and Analysis

Condition	Prompt Length (N)	Token Format	Regression Rate
A (Baseline)	128,000	Raw English History	14.3%
B (English)	8,400	Concise English Rules	9.8%
C (SCP)	1,200	Orthogonal Glyphs (SCP)	**< 0.1%**

Condition A suffered total Channel Capacity Saturation and Posterior Collapse due to unrestricted representation superposition.

Condition B proves that lowering N is necessary but fundamentally insufficient. Despite a 15x reduction in context size relative to the baseline, the prompt still required 8,400 tokens of standard English. Because English natively operates in Strong Superposition, the baseline geometric overlap ($\propto 1/d_k$) of words like "share", "state", and "module" continued to trigger overlapping attention heads, causing inescapable semantic cross-talk. In AI-assisted software engineering, a regression rate close to 10% (9.8%) remains fatal for production codebases.

Comparing Condition B to Condition C definitively isolates the variable of geometric orthogonality. Standard English suffers from interference regardless of compression efforts. By completely removing English semantics, SCP’s 1,200 rare glyphs successfully evaded the Strong Superposition manifold entirely. This artificial injection of Weak Superposition reduced $\bar{E}_{noise}$ to near zero, successfully restoring the $\Delta E > \ln(N)$ gap (Eq. 6), and proving that orthogonal embeddings are strictly necessary to enforce rigid architectural boundaries.

9 Conclusion

The pursuit of the "Billion Token" context window is a pursuit of a thermodynamic impossibility. We cannot solve the Partition Function Explosion simply by adding more memory, because the normalization constant (Z) will always linearly dilute the probability mass of the signal unless the energy gap scales logarithmically ($\Delta E > \ln N$). Furthermore, benchmark evaluations operating on Heterogeneous Noise mask the true geometric interference of Strong Superposition, creating a false sense of security that shatters upon contact with the Homogeneous Noise of production codebases.

"Lost in the Middle" is a symptom; the inflation of Z by the unavoidable $1/d_k$ geometric overlap of adversarial polysemy is the disease. Pouring millions of standard words into the context washes out local constraints, causing AI-assisted software engineering tools to confabulate by relaxing into pre-trained priors. Exhaustive analysis of SOTA variants like SSMs, Linear Attention, and RAG reveals that approaches optimizing computational complexity or positional indexing fail to resolve the underlying geometric interference defined by the Strong Superposition limit. The Spatial Constraint Protocol circumvents this not merely by compressing the prompt, but by utilizing orthogonal mathematical embeddings to escape Strong Superposition entirely, digging *Steeper Valleys* free of interference.

By pairing this with Guided Thermodynamic Forcing via an AST Weaver, we mathematically forbid emergent coupling. To validate models against these real-world limits, the industry must adopt rigorous, exhaustive frameworks like the Enterprise Codebase Regression Benchmark (ECRB) for apples-to-apples evaluation. True autonomous reasoning in AI-assisted software engineering will not emerge from probabilistic engines drowning in infinite, noisy context. It will emerge from Vertical Neuro-Symbolic Integration: pairing the intuitive, generative power of continuous latent spaces with the uncompromising rigor of discrete algorithmic search.

Appendix A: Formal Mathematical Framework Summary

A.1 The Thermodynamics of Attention Decay

- **Standard Attention Mechanism:**

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

- **The Partition Function (Z):**

$$Z = \sum_{j=1}^{N} e^{\text{score}(q, k_j)} \tag{2}$$

A.2 The Critical Energy Gap and Geometric Interference

- **Boltzmann Distribution of Attention:**

$$P_{sig} = \frac{e^{E_{sig}}}{e^{E_{sig}} + \sum_{j \neq sig}^{N} e^{E_{noise,j}}} \tag{3}$$

- **Logarithmic Expansion limit:**

$$E_{sig} > \ln(N) + \bar{E}_{noise} \tag{5}$$

- **Logarithmic Boundary (Critical Energy Gap):**

$$\Delta E = E_{sig} - \bar{E}_{noise} > \ln(N) \tag{6}$$

- **Strong Superposition Welch Bound (Liu et al. 2025):**

$$\max_{i \neq j} |w_i \cdot w_j| \geq \sqrt{\frac{\nu - d_k}{d_k(\nu - 1)}} \approx \sqrt{\frac{1}{d_k}} \tag{7}$$

- **Expected Geometric Overlap:**

$$\mathbb{E}[(w_i \cdot w_j)^2] \propto \frac{1}{d_k} \tag{8}$$

A.3 MAP Instability (Confabulation / Posterior Collapse)

- $$\hat{y} = \arg\max_y P(y|x) \approx \arg\max_y P(y) \tag{9}$$

A.4 The Neuro-Symbolic Resolution (SCP & Weaver)

- **Orthogonal Mapping (Escaping $1/d_k$ Overlap):**

$$\mathbb{E}[\text{sim}(e_{SCP}, e_{distractor})] \approx 0 \tag{10}$$

- **The Weaver Function:**

$$W(G) = \sum_{(i,j) \notin E} MI_{AST}(m_i, m_j) \tag{11}$$

- **Guided Thermodynamic Forcing:**
  If $W(G) > 0$ (Eq. 11) → Extract exact AST node violation → Inject targeted error penalty into prompt X_{t+1} → Resample.

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, such as legacy DB schemas).

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 ($\bar{E}_{noise}$) 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.

References

[1] Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The Long-Document Transformer*.

[2] Casazza, P. G., & Kutyniok, G. (2012). *Finite frames: Theory and applications*. Springer Science & Business Media.

[3] Chevalier, A., et al. (2023). *Adapting Language Models to Compress Contexts*. (AutoCompressors).

[4] Gu, A., & Dao, T. (2023). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*.

[5] Jiang, H., et al. (2023). *LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models*.

[6] Lehman, M. M. (1980). *Programs, Life Cycles, and Laws of Software Evolution*.

[7] Lewis, P., et al. (2020). *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*.

[8] Li, Y., et al. (2024). *The Entropy-Lens Framework*.

[9] Liu, N. F., et al. (2023). *Lost in the Middle: How Language Models Use Long Contexts*.

[10] Liu, Y., Liu, Z., & Gore, J. (2025). *Superposition Yields Robust Neural Scaling*. Massachusetts Institute of Technology. arXiv:2505.10465v4 [cs.LG].

[11] Peng, B., et al. (2023). *RWKV: Reinventing RNNs for the Transformer Era*.

[12] Peng, B., et al. (2023). *YaRN: Efficient Context Window Extension of Large Language Models*.

[13] Reid, M., et al. (2025). *Gemini 1.5 Pro Technical Report*.

[14] Semantic Rate-Distortion Theory (2023).

[15] Su, J., et al. (2024). *RoPE: Rotary Position Embedding*.

[16] Sun, Y., et al. (2023). *RetNet: Retentive Network: A Successor to Transformer for Large Language Models*.

[17] Unified Theory of Latent Space Stability (2024). *Know-But-Don't-Tell Phenomenon*.

[18] Welch, L. (2003). *Lower bounds on the maximum cross correlation of signals*. IEEE Transactions on Information Theory.

[19] Xiao, G., et al. (2024). *Efficient Streaming Language Models with Attention Sinks*.

[20] Zaheer, M., et al. (2020). *Big Bird: Transformers for Longer Sequences*.