The Catenary of Cognition: Why High-Dimensional Attention Naturally Collapses into a U-Shape

Jin Yanyan Independent Researcher Email: lmxxf@hotmail.com

Abstract

The "Lost in the Middle" phenomenon in Large Language Models (LLMs)—where models effectively utilize the beginning and end of long contexts while neglecting the middle—is commonly attributed to architectural limitations or training data bias. This paper proposes a mechanistic explanatory framework: **The Catenary of Cognition**. We argue that in Softmax-dominated attention mechanisms, "semantic tension" is primarily driven by two anchors: **Instruction (Alpha)** and **Query (Omega)**. If middle context lacks sufficient relative scoring advantage (logit margin), its total attention mass will be systematically compressed in normalization competition. This paper provides an explicit upper bound for this compression and presents a minimal mechanistic model demonstrating that when "instruction anchor + recency anchor" coexist, attention distribution can naturally exhibit a U-shape along the positional dimension. The catenary serves in this paper as a physical analogy for energy minimization, explaining "why the U-shape repeatedly appears," rather than claiming that attention curves are strictly equivalent to $y = a\cosh(x/a)$.

---

1. Prologue: The U-Shaped Curse

Research has shown (Liu et al., 2023) that when LLMs face long contexts (e.g., 32k or 128k tokens), their retrieval accuracy is extremely high at the beginning (first 10%) and end (last 10%), but drops significantly in the middle. This "U-shaped" performance curve has long puzzled engineers.

Common engineering explanations include: * **Positional encoding decay:** RoPE or ALiBi attenuates distant signals. * **Training data bias:** Human text habitually sets the tone at the beginning and summarizes at the end. * **Capacity limitations:** Noise interference from KV Cache overload.

While these factors all contribute, they cannot explain the **universality** of this curve—why do models with different architectures, different training data, and different context lengths all exhibit similar U-shapes? We propose that the U-shaped curve is a **topological inevitability** of Transformer attention mechanisms, rooted in the competitive dynamics of Softmax normalization.

---

2. Deconstructing the Geometric Fallacy

An intuitive explanation suggests that in high-dimensional spaces (e.g., $d = 12288$), volume concentrates at the surface while the center is nearly empty (the classic conclusion of the curse of dimensionality; see Bellman 1961). Therefore, middle tokens "fall into the hollow center of the sphere," causing their norm to approach zero.

This is mathematically misleading:

1. **Index vs. Norm:** The "middle" of a text sequence (time $t \approx L/2$) is not the "center" of geometric space (norm $\|v\| \approx 0$).
2. **LayerNorm constraints:** Layer Normalization ensures that regardless of a token's position in the sequence, its vector is firmly distributed on the surface of the hypersphere. The 5000th token's vector length is no different from the 1st token's.

The "middle blind spot" occurs not because tokens geometrically vanish, but because they fail in **topological competition**.

---

3. The Catenary Model: Attention as Tension

We propose the **Catenary Model**. Just as a chain suspended between two poles forms a U-shape due to gravity (a catenary, $y = a\cosh(x/a)$), semantic attention under **Softmax normalization** also sags between two anchors.

3.1 The Two Anchors

Any meaningful generation task is defined by two poles:

1. **Left Anchor: Alpha (Instruction/System Prompt)**
   - Defines the "rules" and "gravitational field" of the semantic universe.
   - It is the "parent node" of all subsequent computation. Attention heads repeatedly look back at it to calibrate output format and task intent.
2. **Right Anchor: Omega (Query/Latest Context)**
   - Defines the "immediacy" of wavefunction collapse.
   - Due to autoregressive nature, the model must pay extreme attention to the most recent tokens to maintain syntactic coherence (recency effect).

3.2 The Middle Sag

The middle context (background documents, conversation history) caught between Alpha and Omega exists in a "tension vacuum":

- **Lack of structural status:** It neither defines rules (Alpha) nor triggers prediction (Omega). It is purely "evidence."
- **Softmax bottleneck:** The Softmax function $\sigma(z)_i = \frac{e^{z_i}}{\sum e^{z_j}}$ is a "winner-take-all" mechanism.
  - **Alpha** scores high due to the authority of its global instructions.
  - **Omega** scores high due to the physical proximity of its position.
  - **The middle** has only mediocre semantic similarity scores. In the denominator's competition, Alpha and Omega's high scores dominate, and the middle's weights are "diluted" to near zero.

Mechanistic conclusion: Under Softmax normalization, as long as Alpha and Omega (or their "anchor clusters") form a stable advantage in logits, the total attention mass of middle tokens will be systematically compressed by denominator competition (see Appendix 8.1). Corpus structure (such as "set tone at beginning, summarize at end") can reinforce this advantage, but it is not the only reason for producing the U-shape; even on synthetic data, as long as "global instruction anchor + local recency anchor" dual bias exists, the U-shape can be produced by mechanism (see Section 7).

---

4. Gradient Starvation

From a training perspective, the U-shaped curve is further solidified through **gradient starvation**:

- **Terminal feedback:** The loss function is computed at the sequence end; gradients flow most directly to the Omega portion.
- **Global accumulation:** The Alpha portion (System Prompt) is attended to by every token during training, accumulating massive gradient updates and becoming a "super node."
- **The forgotten middle:** Middle tokens are only effectively activated in rare "needle-in-a-haystack" cases. On average, the gradient flow they receive is sparse and noisy.

After training on trillions of tokens, the model learns the path of least resistance: **"When in doubt, look at the beginning (find instructions) or look at the recent context (continue the text); scanning the middle is both expensive and uncertain."**

---

5. The Bridge Metaphor and Engineering Implications

Language processing is essentially about building a **semantic bridge**.

- You cannot build a bridge by piling stones in the middle of the river (purely stacking middle context).
- You must build bridge towers on both banks (Alpha & Omega) and suspend the road between them.
- If the span is too long (context too lengthy), the middle will inevitably sag.

Conclusion: To fix "Lost in the Middle," one should not simply "force" the model to attend to the middle (this increases entropy), but rather **add intermediate piers**. For example: hierarchical summarization or introducing "memory anchors" to share the tension load of the catenary through physical support.

---

6. Related Work

Since the Lost in the Middle phenomenon was discovered by Liu et al. (2023), several follow-up studies have emerged:

Engineering remediation: - Found in the Middle (Hsieh et al., 2024) proposes attention calibration methods that improve middle utilization by approximately 15% on RAG tasks through adjusting positional bias. **- Attention Sorting (Peysakhovich & Lerer, 2023)** mitigates recency bias by reordering documents (sorting by attention weights before regenerating).

Mechanistic analysis: - Initial Saliency (Chen et al., 2024) attributes the U-shaped curve to the superposition of "initial token saliency" and "positional encoding bias." **- Limitations of Normalization (Yang et al., 2025)** analyzes the upper bound of token selection capability from the mathematical properties of Softmax normalization.

Unique contributions of this paper: 1. **Catenary analogy:** Uses the energy minimization framework of physical catenaries to explain the intuitive source of "why the U-shape repeatedly appears," while explicitly clarifying that this is an analogy rather than strict isomorphism. 2. **Explicit refutation of the high-dimensional sphere center fallacy:** Points out the confusion between "sequence position" and "vector norm." 3. **Hard inequality for Softmax competition:** Provides an explicit upper bound for middle attention mass (Lemma 8.1), grounding "denominator gravity" from rhetoric to a testable inequality. 4. **Minimal mechanistic model:** Provides a computable dual-anchor positional bias model, deriving sufficient conditions for U-shaped distribution and falsifiable predictions (Section 7).

---

7. Minimal Mechanistic Model & Falsifiability

The goal of this section is to turn "U-shape comes from mechanism" into a statement that can be falsified: provide a minimal model and its falsifiable predictions. If empirical measurements systematically deviate from predictions, the mechanistic narrative of this paper should be downgraded to "an artistic description of corpus priors."

7.1 A Minimal Dual-Anchor Positional Bias Model

In many practical settings, attention logits can be roughly decomposed into two parts: content similarity terms and positional/structural bias terms. To reduce the problem to its minimum, we ignore the details of content terms and retain only "the structural advantage of two anchors":

- Alpha anchor: global saliency from system prompt/task instructions;
- Omega anchor: saliency from autoregressive recency (local window).

Let sequence position be $i \in \{0, 1, \dots, L\}$, with the left anchor at 0 and right anchor at L. Consider the following logit approximation:

$$z_i \;=\; c \;-\; \lambda \cdot \min(i,\, L - i),$$

where $\lambda > 0$ represents "the farther from anchors, the faster structural advantage decays," and c is a constant (can be absorbed into Softmax normalization).

Thus the attention weights are

$$\alpha_i \;\propto\; e^{z_i} \;=\; e^{c}\, e^{-\lambda \min(i, L-i)}.$$

Proposition 7.1 (Dual-anchor exponential decay ⇒ U-shape) If $\lambda > 0$, then α_i strictly decreases with position i on the interval $[0, L/2]$ and strictly increases on the interval $[L/2, L]$; therefore it exhibits a U-shaped distribution along the positional dimension with the valley at the middle and peaks at both ends (when L is even, the valley is at $i = L/2$; when L is odd, the valley is at the adjacent positions $i = \lfloor L/2 \rfloor, \lceil L/2 \rceil$).

Proof: When $0 \le i \le L/2$, $\min(i, L-i) = i$, so $\alpha_i \propto e^{-\lambda i}$, strictly decreasing with i; when $L/2 \le i \le L$, $\min(i, L-i) = L - i$, so $\alpha_i \propto e^{-\lambda(L-i)}$, strictly increasing with i. Q.E.D. □

This model does not claim "real attention is an exponential function"; it only provides a **sufficient condition**: as long as two endpoint anchors exist, and the bias of "farther from anchors is more disadvantageous" holds statistically (whether it comes from RoPE's locality, from accumulated saliency of system prompts during training, or from KV budget and heuristic sparsification), the U-shape is not surprising.

7.2 Mechanistic Explanation vs. Corpus Explanation: How to Distinguish?

To avoid "explaining everything," we decompose competing explanations into two comparable sources:

1. **Mechanistic source (structural bias):** From the competitive nature of Softmax and positional/structural bias;
2. **Corpus source (writing prior):** From the statistical structure of human text where "beginning/end have higher information density."

The two can superimpose, but they respond differently to interventions.

7.3 Falsifiable Predictions (Can Write Acceptance Criteria Without Running Experiments)

Each prediction below corresponds to a "if opposite results are observed, the mechanistic primary cause cannot stand" falsification condition:

1. **Anchor ablation:** If system prompt information is scattered into uniformly distributed local instructions (reducing Alpha anchor), and query is forced to be placed in the middle (weakening Omega recency), the U-shape should significantly flatten; if it doesn't flatten, the U-shape is more likely from corpus structure or other unmodeled biases.
2. **Middle anchor creation:** Insert structured "memory anchors" in the middle (e.g., hierarchical summary titles, retrievable indices, or explicit key-value markers); if only changing structure without changing content can raise middle utilization, it supports the mechanistic narrative that "increasing anchor cluster count K can suppress middle competitive disadvantage" (also consistent with the upper bound intuition in Appendix 8.1).
3. **Reordering invariance:** While keeping "anchor positions and counts unchanged," randomly reorder middle paragraph sequences; if U-shape basically remains but accuracy fluctuates with semantic relevance, it supports "positional bias dominates shape, content similarity dominates details" decomposition; if shape systematically changes with destruction of human writing structure, corpus explanation is stronger.
4. **Genre contrast:** On non-human writing genres or synthetic corpora (uniform information, no paragraph templates), if U-shape remains strong and strongly correlated with "two anchors," it supports mechanistic explanation; if U-shape significantly weakens, it supports corpus explanation.

---

8. Mathematical Appendix: What Do These Two "Proofs" Actually Prove?

This section provides two **testable, reusable** mathematical results to support the "inevitability / minimum energy" phrasing used repeatedly throughout the paper:

1) **Softmax competition leads to "middle weight upper bound"** (independent of implementation details, depending only on score differences);
2) **The classical catenary is the solution for "minimum gravitational potential energy of a uniform chain"** (standard calculus of variations derivation).

These correspond respectively to the "Softmax bottleneck" and "minimum energy shape" statements in the main text.

8.1 Result A: Upper Bound on Total Middle Attention Mass Under Softmax

Consider a single attention head and a single query q's attention distribution. Let the sequence length be L, and the logit at each position i be

$$z_i \;=\; \frac{\langle q, k_i \rangle}{\sqrt{d_k}} \;+\; b_i,$$

where b_i includes positional bias (equivalent effects of RoPE/ALiBi) and any additive structural bias. The attention weights are

$$\alpha_i \;=\; \frac{e^{z_i}}{\sum_{j=1}^{L} e^{z_j}}.$$

We denote the "left anchor" and "right anchor" as two specific positions a, o (Alpha/Omega), and the remaining positions as the set $M = \{1, \dots, L\} \;\; \{a, o\}$ ("middle" here means "non-anchor," not requiring geometric centrality).

Lemma 8.1 (Two-anchor advantage ⇒ middle total weight upper bound) Let

$$m \;=\; \max_{i \in M} z_i.$$

If the anchors satisfy

$$z_a \geq m + \Delta, \qquad z_o \geq m + \Delta$$

for some $\Delta > 0$, then the total attention mass on the "middle"

$$A_M \;=\; \sum_{i \in M} \alpha_i$$

satisfies the upper bound

$$A_M \;\leq\; \frac{(L-2)e^{-\Delta}}{2 + (L-2)e^{-\Delta}}.$$

Proof: By definition, for any $i \in M$ we have $z_i \leq m$, hence

$$\sum_{i \in M} e^{z_i} \leq (L-2)e^{m}.$$

On the other hand, by the anchor advantage condition,

$$e^{z_a} \geq e^{m+\Delta}, \qquad e^{z_o} \geq e^{m+\Delta} \;\Rightarrow\; e^{z_a} + e^{z_o} \geq 2e^{m+\Delta}.$$

Thus

$$A_M = \frac{\sum_{i \in M} e^{z_i}}{e^{z_a} + e^{z_o} + \sum_{i \in M} e^{z_i}} \leq \frac{(L-2)e^{m}}{2e^{m+\Delta} + (L-2)e^{m}} \;=\; \frac{(L-2)e^{-\Delta}}{2 + (L-2)e^{-\Delta}}.$$

Q.E.D. □

Interpretation:

1) This upper bound only uses "anchor logits are Δ higher than the maximum middle logit," thus it characterizes a **structural competition outcome**: as long as Alpha/Omega form a stable advantage in scoring, the middle will be "crushed" by the Softmax denominator.
2) When Δ is fixed and L grows large, the upper bound approaches 1, meaning "two-anchor advantage alone" cannot automatically imply "middle total mass must be small." But in actual attention, what commonly occurs is: anchors are not only higher, but appear as **multi-head, multi-layer, multi-token anchor clusters** (system prompt paragraphs, recent window paragraphs), thus scaling up the "effective anchor count" from 2 to $K \gg 2$, and the upper bound naturally rewrites to

$$A_M \;\le\; \frac{(L-K)e^{-\Delta}}{K+(L-K)e^{-\Delta}},$$

where growth in K significantly suppresses middle total mass.

The significance of this inequality is: the main text's "Softmax winner-take-all" is not mere rhetoric; it can be written as an **explicit bound** on A_M. And one engineering implication of "adding intermediate piers" is to **structure some middle tokens into new anchor clusters** (raising their logits or raising the effective anchor count K).

8.2 Result B: The Catenary Arises from Minimum Gravitational Potential Energy (Classical Calculus of Variations Derivation)

This subsection is unrelated to Transformers; it only answers a pure mathematical question: why does "a uniform chain hanging under gravity" yield $y = a\cosh(x/a)$.

Consider a uniform-density chain with endpoints fixed at (x_1, y_1) and (x_2, y_2), with the y-axis pointing upward and constant gravitational acceleration. Let the chain curve be $y = y(x)$, with arc length element

$$ds = \sqrt{1+y'(x)^2}\,dx.$$

The chain's gravitational potential energy (ignoring constant factors) is proportional to

$$\int y\,ds = \int_{x_1}^{x_2} y(x)\sqrt{1+y'(x)^2}\,dx.$$

The chain length is fixed at S:

$$\int_{x_1}^{x_2} \sqrt{1+y'(x)^2}\,dx = S.$$

Using Lagrange multiplier λ to incorporate the constraint into the functional, this is equivalent to minimizing

$$\mathcal{J}[y] \;=\; \int_{x_1}^{x_2} (y(x)+\lambda)\sqrt{1+y'(x)^2}\,dx.$$

Let

$$F(y,y') = (y+\lambda)\sqrt{1+y'^2}.$$

Note that F does not explicitly contain x, so we can use the Beltrami identity:

$$F - y'\frac{\partial F}{\partial y'} = C$$

for some constant C. Computing the derivative

$$\frac{\partial F}{\partial y'} = (y+\lambda)\frac{y'}{\sqrt{1+y'^2}},$$

therefore

$$F - y'\frac{\partial F}{\partial y'} = \frac{y+\lambda}{\sqrt{1+y'^2}} = C.$$

Rearranging gives

$$\sqrt{1+y'^2} = \frac{y+\lambda}{C} \quad \Rightarrow \quad y'^2 = \left(\frac{y+\lambda}{C}\right)^2 - 1.$$

Taking $a = C$, and letting $u = y + \lambda$, the differential equation becomes

$$\frac{du}{dx} = \pm\sqrt{\left(\frac{u}{a}\right)^2 - 1}.$$

Separating variables:

$$\int \frac{du}{\sqrt{(u/a)^2 - 1}} = \pm \int dx.$$

The left integral gives the inverse hyperbolic cosine:

$$\operatorname{arcosh}\left(\frac{u}{a}\right) = \pm\frac{x - x_0}{a}.$$

Thus

$$u = a\cosh\left(\frac{x - x_0}{a}\right),$$

and substituting back $u = y + \lambda$ gives

$$y(x) = a\cosh\left(\frac{x - x_0}{a}\right) - \lambda,$$

which is the general form of the catenary (constants determined by endpoint and length conditions). Q.E.D. □

8.3 From "Proofs" Back to Main Text: What Are Theorems, What Are Metaphors?

- Section 8.1 provides a **hard inequality for Softmax distribution**: when the two ends (or anchor clusters) have stable advantages in logits, the middle total mass is necessarily compressed; this supports the main text's statement about "gravitational pull of the normalization denominator."
- Section 8.2 provides the **standard minimization theorem for physical catenaries**: the so-called "minimum energy" corresponds to a variational extremum in the strict sense.
- The core claim of the main text is "the attention curve resembles a catenary": strictly speaking, this is a **model analogy**. To elevate the analogy to a theorem requires additionally specifying a precise definition of "attention-energy" (e.g., writing some regularized optimization objective in the form of 8.2). In the current manuscript, we leave this step to future work, while using the hard bound in 8.1 to explain the U-shaped competition mechanism and 8.2 to explain "why catenaries naturally arise."

References

Phenomenon Discovery

1. Liu, N. F., et al. (2023). "Lost in the Middle: How Language Models Use Long Contexts." *TACL 2024*. https://arxiv.org/abs/2307.03172

Engineering Remediation

2. Hsieh, C.-Y., et al. (2024). "Found in the Middle: Calibrating Positional Attention Bias Improves Long Context Utilization." *ACL Findings 2024*. https://arxiv.org/abs/2406.16008
3. Peysakhovich, A. & Lerer, A. (2023). "Attention Sorting Combats Recency Bias in Long Context Language Models." https://arxiv.org/abs/2310.01427

Mechanistic Analysis

4. Chen, Y., et al. (2024). “Uncovering the Role of Initial Saliency in U-Shaped Attention Bias.” https://arxiv.org/abs/2512.13109
5. Yang, Z., et al. (2025). “Limitations of Normalization in Attention Mechanism.” https://arxiv.org/abs/2508.17821

Foundational Architecture

6. Vaswani, A., et al. (2017). “Attention Is All You Need.” *NeurIPS*.

High-Dimensional Geometry Background

7. Bellman, R. (1961). “Adaptive Control Processes: A Guided Tour.” Princeton University Press. (Origin of the curse of dimensionality concept)