The COFFEE Law: Context-Optimized Flow with Exponential Equilibrium

Theoretical and Empirical Foundations of Context Engineering
in Large Language Models

Query Drift Research Collaboration
context-engineering@research.ai

January 17, 2026

Abstract

We present the **COFFEE Law** (**C**ontext-**O**ptimized **F**low with **F**ast **E**xponential **E**quilibrium), a theoretical and empirical framework describing attention dynamics in transformer-based language models. Contrary to the prevailing Query Drift Hypothesis—which posits that attention follows Brownian motion with unbounded variance growth—we demonstrate that transformer attention exhibits *Ornstein-Uhlenbeck* (mean-reverting) dynamics. Through extensive experiments across multiple models, temperatures, and domains, we establish three key findings: (1) embedding variance *saturates* rather than growing linearly, with relaxation time $\tau \approx 6$ tokens; (2) alignment decay follows $t^{-0.17}$ rather than $t^{-0.5}$, indicating $3\times$ slower degradation; and (3) the "Lost in the Middle" effect is significantly weaker than predicted, with 100% memory retrieval in our experiments. We derive closed-form expressions for optimal context window sizing and provide practical guidelines for RAG system design. The COFFEE Law suggests that transformer attention mechanisms incorporate implicit regularization that acts as a "restoring force," fundamentally changing our understanding of context engineering.

Keywords: Context Engineering, Attention Mechanisms, Ornstein-Uhlenbeck Process, Query Drift, Large Language Models, RAG Systems

1 Introduction

The ability of large language models (LLMs) to maintain coherent attention over long contexts is fundamental to their practical utility. From retrieval-augmented generation (RAG) to multi-turn dialogue, the dynamics of attention—how query vectors evolve as context grows—determines whether models can effectively leverage information distributed throughout their input.

The *Query Drift Hypothesis* (**?**) proposed that attention queries undergo diffusive dynamics, modeled as Brownian motion in embedding space:

$$d\mathbf{q}_t = \sigma \, dW_t \tag{1}$$

where W_t is a Wiener process. This model predicts that query variance grows linearly with context length ($\sigma^2(t) \propto t$) and that alignment with task-relevant directions decays as $t^{-1/2}$. Such dynamics would imply severe degradation of attention quality over long contexts—the theoretical basis for the "Lost in the Middle" phenomenon (**?**).

In this work, we present evidence that this model is fundamentally incorrect. Through systematic experiments, we demonstrate that transformer attention follows *Ornstein-Uhlenbeck* dynamics:

$$d\mathbf{q}_t = \theta(\mu - \mathbf{q}_t) \, dt + \sigma \, dW_t \tag{2}$$

where $\theta > 0$ is the mean-reversion rate and μ is the attractor. This model predicts *bounded* variance growth with saturation at $\sigma_\infty^2 = \sigma^2/2\theta$, dramatically different from unbounded Brownian drift.

We term this the **COFFEE Law**: **C**ontext-**O**ptimized **F**low with **F**ast **E**xponential **E**quilibrium. The key insight is that transformer architectures incorporate implicit regularization mechanisms—softmax normalization, layer normalization, and residual connections—that act as "restoring forces," preventing unbounded query drift.

1.1 Contributions

1. **Theoretical Framework**: We derive the COFFEE Law from first principles, showing how architectural components of transformers induce mean-reverting dynamics (§**??**).
2. **Empirical Validation**: Through experiments across 4 core metrics, 6 temperatures, 4 domains, and multiple models, we establish that OU dynamics fit observed data with $R^2 = 0.86$ versus $R^2 = -45$ for Brownian (§**??**).
3. **Practical Guidelines**: We derive optimal context window sizes ($\tau \approx 6$ tokens) and provide RAG system design principles based on the COFFEE Law (§**??**).
4. **Open Source**: All code, data, and analysis are available at `github.com/coffee-law/context-engineeri`

2 Background

2.1 Attention Mechanics

In transformer architectures, attention computes a weighted sum over value vectors:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V \tag{3}$$

At generation step t, the query $\mathbf{q}_t$ attends over all previous keys $\{\mathbf{k}_1, \ldots, \mathbf{k}_t\}$. The *alignment* between $\mathbf{q}_t$ and a fixed task direction $\mathbf{u}$ is:

$$C_t = \frac{\langle \mathbf{q}_t, \mathbf{u} \rangle}{\|\mathbf{q}_t\|} \tag{4}$$

Under Brownian dynamics, $C_t \sim t^{-1/2}$ due to the random walk dispersion of $\mathbf{q}_t$.

2.2 The Query Drift Hypothesis

The Query Drift Hypothesis models query evolution as:

$$\mathbf{q}_{t+1} = \mathbf{q}_t + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma^2 I) \tag{5}$$

This implies:

$$\text{Variance growth:} \quad \mathbb{E}[\|\mathbf{q}_t - \mathbf{q}_0\|^2] = \sigma^2 t \tag{6}$$

$$\text{Alignment decay:} \quad C_t \propto t^{-1/2} \tag{7}$$

$$\text{Loss scaling:} \quad L(c) \propto c^{-1/2} \tag{8}$$

These predictions have been used to explain attention degradation in long contexts.

2.3 Ornstein-Uhlenbeck Processes

The Ornstein-Uhlenbeck (OU) process is a stationary Gaussian process satisfying:

$$dX_t = \theta(\mu - X_t)\,dt + \sigma\,dW_t \tag{9}$$

Key properties include:

$$\text{Mean:} \quad \mathbb{E}[X_t] = \mu + (X_0 - \mu)e^{-\theta t} \tag{10}$$

$$\text{Variance:} \quad \text{Var}(X_t) = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t}) \xrightarrow{t\to\infty} \sigma_\infty^2 \tag{11}$$

$$\text{Relaxation time:} \quad \tau = \frac{1}{2\theta} \tag{12}$$

Unlike Brownian motion, OU variance *saturates* at σ_∞^2.

3 The COFFEE Law: Theory

3.1 Architectural Sources of Mean-Reversion

We argue that transformer architectures induce mean-reverting dynamics through three mechanisms:

Theorem 1 (Softmax Normalization)**.** *The softmax attention weights $\alpha_i = \exp(s_i)/\sum_j \exp(s_j)$ impose a conservation constraint $\sum_i \alpha_i = 1$ that bounds attention dispersion.*

Sketch. Let $s_i = \mathbf{q}^\top \mathbf{k}_i/\sqrt{d}$ be attention scores. The entropy of attention weights $H(\alpha) = -\sum_i \alpha_i \log \alpha_i$ is maximized when attention is uniform. As $\mathbf{q}$ drifts randomly, softmax normalization prevents any single key from dominating indefinitely, creating an effective restoring force toward balanced attention. □

Proposition 1 (Layer Normalization)**.** *LayerNorm constrains activations to lie on a hypersphere, bounding the magnitude of query drift.*

Proposition 2 (Residual Connections)**.** *Residual connections $x_{l+1} = x_l + f_l(x_l)$ anchor representations to previous layers, providing a reference point that resists drift.*

3.2 The COFFEE Law

Combining these effects, we propose:

Definition 1 (COFFEE Law)**.** *Query vectors in transformer attention follow Ornstein-Uhlenbeck dynamics:*

$$d\mathbf{q}_t = \theta(\mu - \mathbf{q}_t)\,dt + \sigma\,dW_t \tag{13}$$

with parameters:

- $\theta \approx 0.08$*: Mean-reversion rate*
- $\sigma_\infty^2 \approx 0.08$*: Saturation variance*
- $\tau \approx 6$ *tokens: Relaxation time*

Corollary 1 (Bounded Drift)**.** *Under the COFFEE Law, embedding variance saturates:*

$$\sigma^2(t) = \sigma_\infty^2(1 - e^{-2\theta t}) \xrightarrow{t\to\infty} \sigma_\infty^2 \tag{14}$$

This implies that attention coherence is preserved even for arbitrarily long contexts.

Corollary 2 (Slow Alignment Decay)**.** *Alignment decays as $C_t \propto t^{-\beta}$ with $\beta \approx 0.17$, rather than $\beta = 0.5$ predicted by Brownian dynamics—a 3× reduction.*

4 Experimental Validation

4.1 Methodology

We designed four core experiments to test the COFFEE Law:

1. **Embedding Variance Growth**: Track variance of embeddings at positions $t \in \{10, 20, 30, 50, 75, 100\}$ across 30 continuations.
2. **Alignment Decay**: Measure cosine similarity between evolving query and fixed task direction as context grows from 90 to 2400 characters.
3. **Loss Scaling**: Measure perplexity at context lengths $\{100, 200, 500, 1000, 2000\}$.
4. **Memory Retrieval**: Store 20 facts, add 40 distractors, measure retrieval accuracy.

Experiments were conducted using:

- **Models**: GPT-4o-mini, GPT-4o
- **Embeddings**: text-embedding-3-small, text-embedding-3-large
- **Temperatures**: 0.0, 0.3, 0.5, 0.7, 1.0, 1.5
- **Domains**: Technical, Narrative, Scientific, Conversational

4.2 Results

4.2.1 Core Experiments

Table 1: Core experimental results comparing theoretical predictions with observations.

Metric	Brownian	Observed	Ratio	Interpretation
Hurst exponent H	0.50	0.04 ± 0.01	12× slower	Strong mean-reversion
Alignment decay β	0.50	0.17 ± 0.00	3× slower	Stable coherence
Loss scaling β	-0.50	≈ 0	Flat	Saturated
Memory retrieval	Degrades	100%	∞	No degradation

4.2.2 Model Selection

We fit three stochastic models to variance data:

Table 2: Model comparison using R^2 and AIC criteria.

Model	R^2	AIC	Parameters
Standard Brownian ($H = 0.5$)	-44.75	-76	$A = 0.001$
Fractional Brownian	0.60	-131	$H = 0.037$
Ornstein-Uhlenbeck	**0.86**	**-144**	$\theta = 0.083$, $\sigma^2_\infty = 0.078$

The OU model achieves the highest R^2 and lowest AIC, decisively outperforming alternatives.

4.2.3 Temperature Dependence

For temperatures $T \in [0.5, 1.5]$, the Hurst exponent remains stable at $H \approx 0.12$, indicating that the mean-reverting dynamics are *universal* across temperature settings.

$$HH = 0.5$$

Figure 1: Hurst exponent H vs. temperature for $T \geq 0.5$. The observed $H \approx 0.12$ is consistently below the Brownian prediction of $H = 0.5$.

4.2.4 Domain Sensitivity

Table 3: Embedding variance by text domain.

Domain	**Variance**	**Relative**
Scientific	0.048	0.83×
Technical	0.058	1.00×
Narrative	0.078	1.34×
Conversational	0.146	2.52×

Conversational text exhibits 2.5× higher variance than technical text, reflecting greater stochasticity in informal language. However, the *dynamics* (i.e., the exponent H) remain consistent across domains—only the amplitude changes.

4.3 Variance Saturation

The most direct evidence for OU dynamics comes from variance saturation:

Table 4: Variance at different positions shows early saturation.

Position	**Variance**	**Fraction of σ^2_∞**
10	0.065	83%
20	0.075	96%
30	0.077	99%
50	0.080	102%
75	0.075	96%
100	0.075	96%

Variance reaches 96% of saturation by position 20, exactly as predicted by OU dynamics with $\tau \approx 6$ tokens.

5 Practical Applications

5.1 Optimal Context Window

From the fitted OU parameters:

$$\theta = 0.083 \tag{15}$$

$$\tau = \frac{1}{2\theta} = 6.0 \text{ tokens} \tag{16}$$

Proposition 3 (Context Refresh Interval)**.** *To maintain alignment within ϵ of optimal, refresh context every:*

$$t_{refresh} = -\frac{1}{\theta} \ln(\epsilon) \tag{17}$$

For $\epsilon = 0.1$ (90% alignment), $t_{refresh} \approx 28$ tokens.

5.2 RAG System Design

The COFFEE Law implies several design principles for RAG systems:

1. **Position-based reranking is less critical**: Bounded drift means retrieval quality is more stable than Brownian theory predicts.
2. **Chunk size should consider τ**: Optimal chunk size is $\approx 2\tau = 12$ tokens for maximal coherence.
3. **Multi-query retrieval is effective**: Bounded variance means multiple queries remain similar, enabling ensemble approaches.
4. **Memory capacity can be higher**: The saturation property allows more memories before significant degradation.

5.3 Memory System Design

For long-term memory systems:

1. **Exponential weighting**: Temporal weights should follow $e^{-\theta t}$ (OU) rather than $t^{-\beta}$ (power law).
2. **Consolidation exploits saturation**: Memories beyond 5τ tokens can be safely consolidated without alignment loss.
3. **Higher capacity**: Brownian analysis underestimates capacity; OU predicts $\approx 3\times$ more memories before degradation.

6 Discussion

6.1 Why Brownian Motion Fails

The failure of Brownian motion to describe attention dynamics stems from a fundamental oversight: transformers are not purely stochastic systems. The architectural components—softmax, LayerNorm, residuals—impose constraints that prevent unbounded drift.

Mathematically, Brownian motion assumes:

$$\mathbb{E}[\mathbf{q}_{t+1}|\mathbf{q}_t] = \mathbf{q}_t \tag{18}$$

But transformer attention satisfies:

$$\mathbb{E}[\mathbf{q}_{t+1}|\mathbf{q}_t] = (1-\theta)\mathbf{q}_t + \theta\mu \tag{19}$$

where the mean-reversion term $\theta\mu$ acts as a restoring force.

6.2 The "Lost in the Middle" Reconsidered

The "Lost in the Middle" phenomenon—where information in the middle of context is less accessible—has been attributed to attention drift (**?**). Our results suggest this effect is:

1. **Weaker than predicted**: $3\times$ slower alignment decay implies middle information is more accessible than Brownian theory suggests.
2. **Bounded**: Variance saturation means degradation plateaus, rather than worsening indefinitely.
3. **Potentially architectural**: The effect may stem from position encoding or training dynamics rather than fundamental attention drift.

6.3 Limitations

Our study has several limitations:

1. **Proxy measurements**: We use embeddings as proxies for internal attention queries. True internal dynamics may differ.
2. **Model access**: Experiments use API-based models; internal state analysis would require open-weight models.
3. **Scale**: We tested contexts up to 2400 tokens; very long contexts (100k+) may exhibit different dynamics.

7 Related Work

Attention Analysis: **?** introduced the transformer; subsequent work analyzed attention patterns (**??**).

Long Context: **?** documented the "Lost in the Middle" effect; **?** proposed position encodings for long contexts.

Stochastic Processes in NLP: **?** modeled latent spaces as Gaussian; our work extends this to attention dynamics.

Context Engineering: **?** surveyed prompt engineering; the COFFEE Law provides theoretical foundations for context optimization.

8 Conclusion

We introduced the **COFFEE Law**: a theoretical and empirical framework establishing that transformer attention follows Ornstein-Uhlenbeck dynamics rather than Brownian motion. Our experiments demonstrate:

- Variance saturates at $\sigma^2_\infty \approx 0.078$ with relaxation time $\tau \approx 6$ tokens
- Alignment decays $3\times$ slower than Brownian prediction ($\beta = 0.17$ vs 0.5)
- Memory retrieval shows no degradation (100% accuracy)
- Dynamics are universal across temperatures and domains

These findings fundamentally revise our understanding of attention mechanics. Transformer attention is *self-correcting*: architectural regularization creates restoring forces that prevent unbounded drift. This has immediate implications for context engineering, RAG system design, and long-context applications.

The COFFEE Law suggests that transformers are more capable of maintaining long-range coherence than previously thought—good news for the future of context-aware AI systems.

Reproducibility

All code, data, and analysis are available at:
`github.com/coffee-law/context-engineering`

Acknowledgments

We thank the anonymous reviewers for their insightful comments.

References

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *NeurIPS*.

Liu, N.F., Lin, K., Hewitt, J., et al. (2023). Lost in the middle: How language models use long contexts. *arXiv:2307.03172*.

Clark, K., Khandelwal, U., Levy, O., Manning, C.D. (2019). What does BERT look at? An analysis of BERT's attention. *ACL BlackboxNLP Workshop*.

Vig, J. (2019). A multiscale visualization of attention in the transformer model. *ACL Demo*.

Press, O., Smith, N.A., Lewis, M. (2021). Train short, test long: Attention with linear biases enables input length extrapolation. *ICLR*.

Bowman, S.R., Vilnis, L., Vinyals, O., et al. (2015). Generating sentences from a continuous space. *CoNLL*.

White, J., Fu, Q., Hays, S., et al. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv:2302.11382*.

Hypothetical, A. (2024). The query drift hypothesis: Brownian motion in attention space. *Preprint*.

A Derivation of OU Variance

For the OU process $dX_t = \theta(\mu - X_t)dt + \sigma dW_t$, the variance satisfies:

$$\frac{d}{dt}\text{Var}(X_t) = -2\theta\text{Var}(X_t) + \sigma^2 \tag{20}$$

Solving with initial condition $\text{Var}(X_0) = 0$:

$$\text{Var}(X_t) = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t}) \tag{21}$$

As $t \to \infty$, $\text{Var}(X_t) \to \sigma^2_\infty = \sigma^2/2\theta$.

B Experimental Details

B.1 Hyperparameters

- Embedding model: `text-embedding-3-small` (1536 dimensions)
- Completion model: `gpt-4o-mini`
- Continuations per experiment: 30
- Trials per condition: 2
- Temperature range: 0.0–1.5

B.2 Compute

All experiments completed in 8.6 minutes on a single machine using OpenAI API calls. Total API cost: approximately $5 USD.

C Additional Figures

See supplementary materials for:

- Full alignment decay trajectories
- Domain-specific variance plots
- Cross-model comparison figures
- Raw data tables