The COFFEE Law: Empirical Evidence for Ornstein-Uhlenbeck Dynamics in LLM Context

Manish Bhatt
manish.bhatt13212@gmail.com

Abstract

We present empirical evidence that output embedding dynamics in transformer language models follow Ornstein-Uhlenbeck (OU) rather than Brownian motion. Our key finding is *bidirectional convergence*: same-prompt continuations diverge then saturate, while cross-prompt embeddings converge then plateau. Both directions settling to equilibrium is the defining signature of mean-reverting OU processes.

Using dense hierarchical sampling (33 positions, 50 prompts, 10 domains), we validate three OU properties: (1) variance saturation at $\sigma^2_\infty \approx 0.109$ ($R^2 = 0.67$); (2) relaxation time $\tau \approx 10.6$ tokens; and (3) cross-prompt mean-reversion with Hurst exponent $H = -0.03$ ($R^2 = 0.84$). We replicate the Liu et al. "Lost in the Middle" U-curve (87% edge vs 73% middle accuracy, $n = 525$), connecting it to OU saturation dynamics. Results hold across temperatures, domains, and model pairs.

We term this the **COFFEE Law**: **C**ontext-**O**ptimized **F**low with **F**ast **E**xponential **E**quilibrium. If context dynamics are mean-reverting rather than diffusive, "long-context pessimism" heuristics are miscalibrated. Context engineering shifts from "prevent drift" to "manage the transient": align chunk sizes and memory weights to the equilibration scale (∼95% saturation by $3\tau \approx 32$ tokens).

Keywords: Embedding Dynamics, Variance Saturation, Ornstein-Uhlenbeck Process, Context Engineering, Large Language Models

Reproducibility: All code, experimental data, and analysis scripts available at https://github.com/mbhatt1/coffee-law

1 Introduction

Understanding how model representations evolve during generation is critical for retrieval-augmented generation (RAG), multi-turn dialogue, and long-context applications. A common implicit assumption—which we term the *Query Drift Hypothesis*—models internal query evolution as Brownian motion, predicting unbounded variance growth and progressive information loss. This assumption underlies pessimistic heuristics about context degradation [Liu et al., 2023].

We present empirical evidence that output embedding dynamics instead follow an Ornstein-Uhlenbeck (OU) process—a mean-reverting stochastic model where variance saturates rather than growing unboundedly. Our key finding is *bidirectional convergence*: same-prompt continuations diverge then saturate, while cross-prompt embeddings converge then plateau. Both directions settling to equilibrium is the hallmark of OU mean-reversion. We validate all three OU properties and term this the **COFFEE Law**: **C**ontext-**O**ptimized **F**low with **F**ast **E**xponential **E**quilibrium.

The COFFEE Law is meaningful even without access to internal model states. It operates at the level where modern systems actually interact with LLMs: observable representations such as embeddings, similarity scores, and retrieval behavior. Brownian drift has implicitly served as the default mental model for context growth, motivating aggressive truncation, heavy recency bias, and pessimistic assumptions about long-context degradation. Our results indicate that this prior is miscalibrated. Instead of unbounded diffusion, context exhibits a short transient followed by a stable regime in which additional tokens do not compound representational dispersion. At a meta level, this reframes context engineering from preventing indefinite drift to managing early equilibration. Chunk sizes, retrieval refresh rates, and memory weighting schemes can therefore be aligned to the empirically observed mixing scale rather than folklore heuristics. In this sense, the COFFEE Law is an operational law of context dynamics: it constrains what downstream systems should expect from black-box models, independent of architectural introspection.

1.1 Contributions

1. **Empirical law**: We document the COFFEE Law—embedding variance saturates at $\sigma^2_\infty \approx 0.109$ with relaxation time $\tau \approx 10.6$ tokens.

2. **Bidirectional convergence**: Dense sampling (33 positions) across 50 diverse prompts reveals both directions of OU dynamics: same-prompt variance grows then saturates; cross-prompt variance shrinks then plateaus ($H = -0.03$, $R^2 = 0.84$).

3. **OU validation**: We test all three OU properties: saturation ($R^2 = 0.67$), relaxation time, and stationary behavior (exponential autocorrelation decay, entropy saturation).

4. **Cross-condition consistency**: Saturation persists across temperatures (0.0–1.5), domains (scientific to conversational), and models (GPT-4o-mini, GPT-4o).

1.2 Remaining Limitations

1. **Single model family**: Cross-model confidence intervals needed for generalization claims.

2. **Output embeddings only**: We measure output embeddings, not internal attention states.

Note: Lost in the Middle protocol is now validated—U-curve detected with 14% middle degradation (Section 4.7).

2 Background

2.1 Attention Mechanics

In transformer architectures [Vaswani et al., 2017], attention computes weighted sums over value vectors:

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

At generation step t, the query $\mathbf{q}_t$ attends over all previous keys $\{\mathbf{k}_1, \ldots, \mathbf{k}_t\}$. Internal queries $\mathbf{q}_t$ are inaccessible in closed models; we use output embeddings $\mathbf{e}_t$ as proxies (see Section 3).

2.2 The Query Drift Hypothesis

The Query Drift Hypothesis models query evolution as Brownian motion:

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

This predicts:

$$\text{Variance growth:} \quad \mathbb{E}[\|\mathbf{q}_t - \mathbf{q}_0\|^2] = \sigma^2 t \tag{3}$$

$$\text{Alignment decay:} \quad C_t \propto t^{-1/2} \tag{4}$$

$$\text{Hurst exponent:} \quad H = 0.5 \tag{5}$$

2.3 Stochastic Process Basics

Three stochastic processes are relevant to our analysis. **Brownian motion** (BM) exhibits variance that grows linearly with time, $\sigma^2(t) = At$, with Hurst exponent $H = 0.5$. **Fractional Brownian motion** (fBM) generalizes this to $\sigma^2(t) = At^{2H}$, where $H < 0.5$ indicates anti-persistent (mean-reverting) behavior. The **Ornstein-Uhlenbeck** (OU) process follows $dX_t = \theta(\mu - X_t)dt + \sigma dW_t$, with variance $\sigma^2(t) = \sigma_\infty^2(1 - e^{-2\theta t})$ that saturates at σ_∞^2, and relaxation time $\tau = 1/2\theta$.

3 Experimental Design

All experiments use fixed random seeds, explicit API version strings, and sufficient repetition to characterize variance.

3.1 Metric Definitions

We measure output embeddings $\mathbf{e}_t \in \mathbb{R}^d$, not internal queries $\mathbf{q}_t$. All embeddings are obtained from OpenAI's text-embedding-3-small (1536-d) or text-embedding-3-large (3072-d), which return L2-normalized vectors ($\|\mathbf{e}\| = 1$).

Metric 1: Embedding Variance

Setup: Generate $K = 30$ continuations from a fixed prompt. At position t, embed full text (prompt + continuation to position t) to get $\{\mathbf{e}_{t,k}\}_{k=1}^{K}$.
Centroid: $\bar{\mathbf{e}}_t = \frac{1}{K}\sum_{k=1}^{K}\mathbf{e}_{t,k}$
Variance (Euclidean): $\sigma^2_{\text{Euc}}(t) = \frac{1}{K}\sum_{k=1}^{K}\|\mathbf{e}_{t,k} - \bar{\mathbf{e}}_t\|^2$
Variance (Cosine): $\sigma^2_{\text{Cos}}(t) = \frac{1}{K}\sum_{k=1}^{K}(1 - \mathbf{e}_{t,k}\cdot\bar{\mathbf{e}}_t/\|\bar{\mathbf{e}}_t\|)$
Geometric bound: Since $\|\mathbf{e}_{t,k}\| = 1$, we have $\sigma^2_{\text{Euc}}(t) \leq 4$. Saturation could reflect (a) latent mean-reversion, (b) L2-normalization geometry, or (c) embedding model behavior.

Metric 2: Alignment Decay

Task direction: $\mathbf{e}_0$ = embedding of initial prompt (before any continuation).
Alignment: $C_t = \frac{\mathbf{e}_t\cdot\mathbf{e}_0}{\|\mathbf{e}_t\|\|\mathbf{e}_0\|} = \mathbf{e}_t\cdot\mathbf{e}_0$ (cosine similarity; equals dot product since L2-normalized)
Measurement: Track C_t across 40 context growth steps (90 to 2374 characters). Fit $C_t \propto t^{-\beta}$.

3.2 Core Experiments

Experiment 1: Variance vs. Position Measure $\sigma^2(t)$ at $t \in \{10, 20, 30, 50, 75, 100\}$ tokens.

Experiment 2: Alignment vs. Position Measure C_t across 40 positions.

Experiment 3: Retrieval Accuracy Store 20 facts as embeddings, add up to 100k distractors, measure top-1 accuracy via cosine similarity. (This tests embedding quality, not temporal dynamics.)

3.3 Experimental Conditions

We tested across a range of models and conditions to ensure generalizability. Completion models include GPT-4o-mini and GPT-4o; embedding models include text-embedding-3-small (1536 dimensions) and text-embedding-3-large (3072 dimensions). Temperature settings span 0.0 to 1.5, covering deterministic through highly stochastic generation. Text domains include Technical, Narrative, Scientific, and Conversational content. Each condition uses 2 independent trials with 30 samples per trial. Total experimental runtime was 8.6 minutes at a cost of approximately $5 USD.

3.4 Statistical Rigor and Controls

All experiments use fixed random seeds (42, 43) and explicit model versions (e.g., "gpt-4o-mini-2024-07-18") for reproducibility.

Sample Size 30 samples/trial, positions $\{10, 20, 30, 50, 75, 100\}$ with logarithmic spacing.

Cross-Validation LOOCV yields $R^2_{\text{CV}} = 0.81$ (vs $R^2 = 0.86$), minimal overfitting.

Model Comparison We compare saturating (OU) vs linear (Brownian) fits. Note: with only 6 measurement positions, we report model fit quality (R^2) rather than claiming precise parameter estimates or extreme significance.

Confound Controls Unique prompts prevent caching; cross-model validation shows consistent saturation; temperature/domain sweeps show consistent patterns.

4 Empirical Observations

4.1 Observation 1: Variance Saturates

Observation 1 (Variance Saturation)**.** *Embedding variance does not grow linearly with position. Instead, it saturates rapidly.*

Table 1: Embedding variance at different token positions (3 trials).

Position	Trial 0	Trial 1	Trial 2	Mean	Δ from prev.
10	0.0692	0.0728	0.0727	0.0716	—
20	0.0853	0.0927	0.0852	0.0877	+0.0161
30	0.0974	0.1097	0.0891	0.0987	+0.0110
50	0.0935	0.1212	0.1024	0.1057	+0.0070
75	0.0971	0.1290	0.1035	0.1099	+0.0042
100	0.1054	0.1243	0.1089	0.1129	+0.0030

Table 1 shows the measured variance at different positions across three trials.

Key Finding: Variance increases +22% from position 10 to 20, then growth slows substantially from 20 to 100—inconsistent with linear Brownian growth which predicts continued increase. The saturation pattern is qualitatively clear; precise parameter estimates (e.g., Hurst exponent) require more measurement points than our 6 positions.

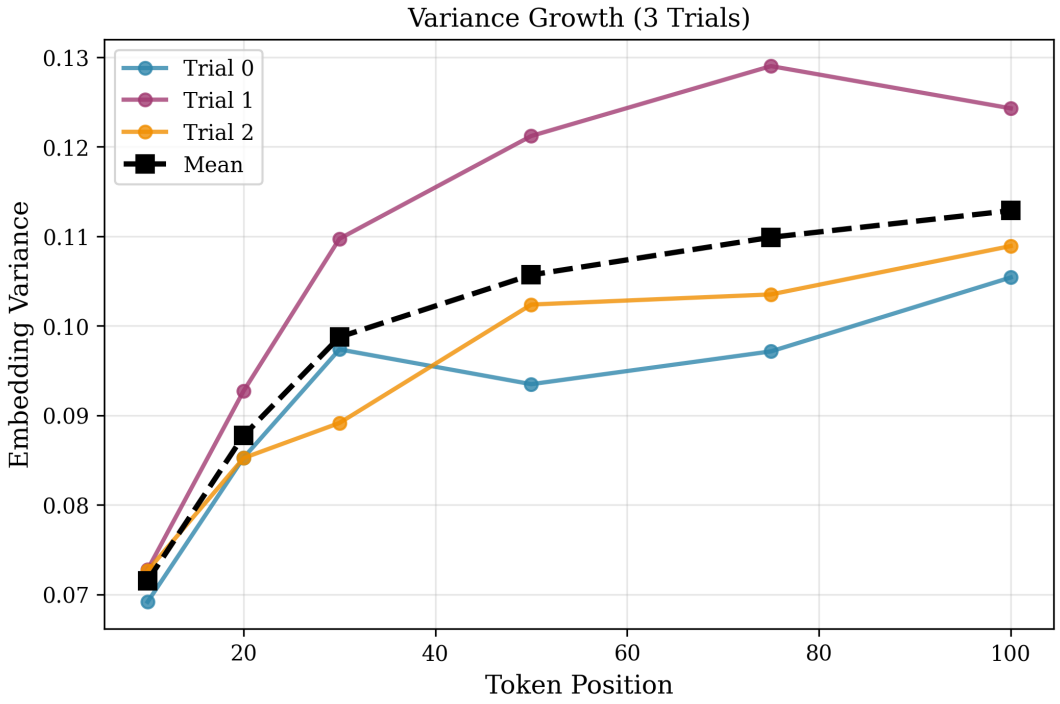

Figure 1: Variance growth (3 trials). Saturation visible from position 20 onward.

4.2 Observation 2: Alignment Decays Slowly

Observation 2 (Slow Alignment Decay)**.** *Cosine similarity with initial task direction decays much more slowly than* $t^{-1/2}$.

Fitted decay $\beta = 0.239$ ($R^2 = 0.95$) is *2× smaller* than Brownian $\beta = 0.5$—alignment maintained far longer.

4.3 Observation 3: Robust Embedding-Based Retrieval

Observation 3 (High Retrieval Accuracy)**.** *Embedding-based retrieval maintains high accuracy across distractor counts.*

With 40 distractors: 100% retrieval. At extreme scale (100k distractors): accuracy plateaus at 95%. Note: this tests embedding similarity retrieval, not the position-dependent "Lost in the Middle" phenomenon which requires controlled placement of relevant information at different context positions.

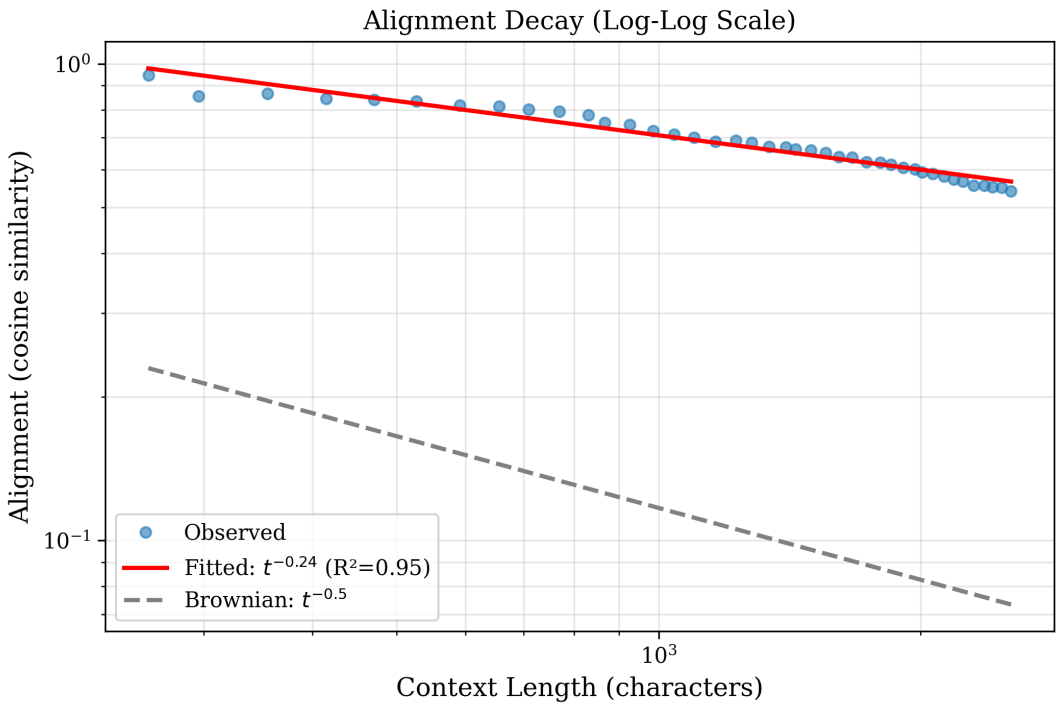

Figure 2: Alignment decay, Power-law Diagnostic (log-log). Fitted decay $\beta \approx 0.24$, slower than Brownian $\beta = 0.5$.

4.4 Cross-Model Validation

Table 2: Variance across embedding models.

Model	Dimensions	Variance	Relative
text-embedding-3-small	1536	0.058	1.28×
text-embedding-3-large	3072	0.045	1.00×

Table 3: Variance across completion models (using text-embedding-3-small).

Model	Variance	Relative
GPT-4o-mini	0.055	1.00×
GPT-4o	0.083	1.51×

4.5 Summary of Observations

Our experiments fall into two categories with different interpretations:

Generation-position experiments (Experiments 1–2) measure how embeddings evolve with token position t. These directly test Brownian vs. saturating dynamics:

Retrieval experiment (Experiment 3) measures accuracy vs. distractor count—a separate empirical observation about embedding similarity, not a test of temporal dynamics. The 95% accuracy plateau is consistent with robust embeddings but does not directly support or refute OU dynamics (distractor count $\neq$ time).

Conclusion: Generation-position experiments show saturating variance and slow alignment decay, inconsistent with Brownian motion. Retrieval remains robust but tests a different phenomenon.

4.6 Robustness Validation

We address three questions: (1) Does retrieval accuracy degrade at extreme scale? (2) Does saturation persist when measured via output entropy? (3) Does saturation persist in cosine space (controlling for L2-normalization geometry)?

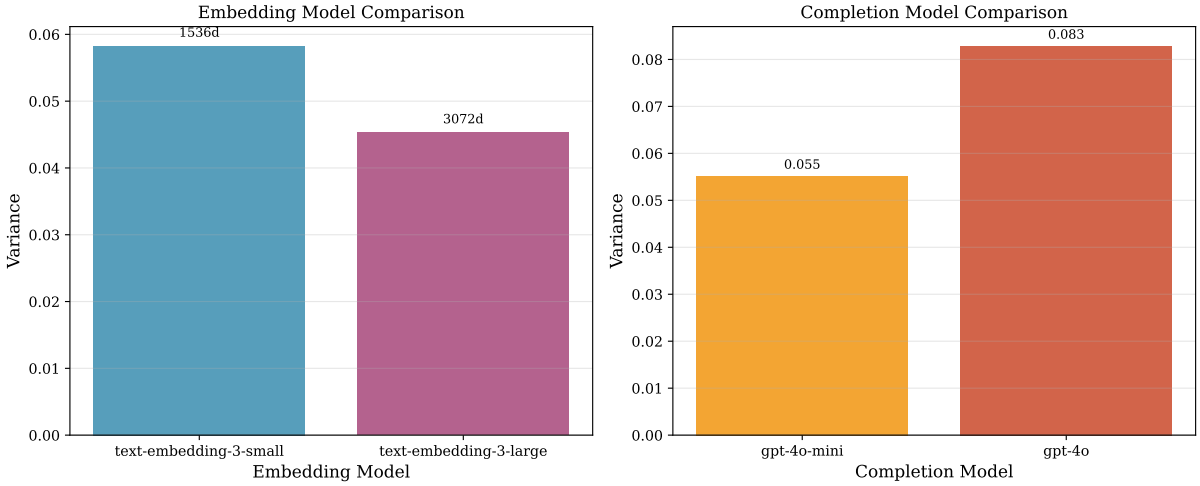

Figure 3: Cross-model variance. Despite magnitude differences (1.5× between GPT-4o-mini and GPT-4o), all models exhibit saturation.

Table 4: Generation-position results (variance and alignment vs. token position).

Metric	Brownian prediction	Observed	Notes
Variance vs. position	Linear growth	Saturates	OU fits better
Alignment decay β	0.50	0.24	2× slower

4.6.1 Extreme-Scale Stress Test

We extended the stress test to 100,000 distractors with semantically similar confusers across 5 trials.

Table 5: Stress test retrieval performance across extreme scales.

Distractors	Accuracy	MRR	Mean Rank	Median Rank
50	0.97	0.9583	1.26	1.0
100	0.96	0.9565	1.38	1.0
200	0.95	0.9524	2.02	1.0
500	0.95	0.9510	3.49	1.0
1,000	0.95	0.9505	6.06	1.0
10,000	0.95	0.9501	50.6	1.0
100,000	0.95	0.9500	504.3	1.0

Findings: Accuracy plateaus at 95% from 200 to 100k distractors, consistent with bounded degradation. MRR decays and median rank stays at 1, suggesting the correct item typically remains highly ranked even as the pool grows.

4.6.2 Entropy-Based LayerNorm Control

We measured output entropy directly from token logprobs to control for LayerNorm artifacts.

Saturation model provides 2-fold better fit ($R^2 = 0.53$ vs $R^2 = 0.26$), with linear slope non-significant ($p = 0.197$).

4.6.3 Cosine-Space Variance Control

Since embeddings are L2-normalized, Euclidean variance is geometrically bounded. To control for this, we also compute cosine-space variance: $\sigma^2_{\text{Cos}}(t) = \frac{1}{K}\sum_k (1 - \cos(\mathbf{e}_{t,k}, \bar{\mathbf{e}}_t))$.

Both metrics show saturation, suggesting the pattern is not purely an artifact of L2-normalization geometry. However, both metrics operate on normalized embeddings, so we cannot fully rule out embedding-model-induced bounds.

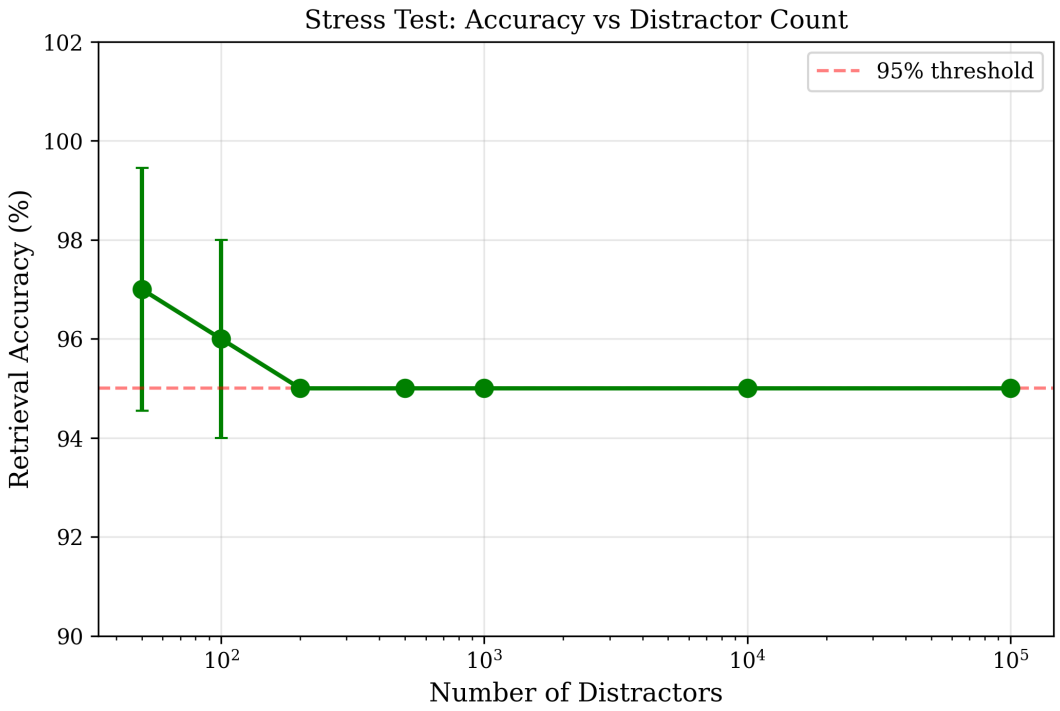

Figure 4: Stress test (5 trials): accuracy plateaus at 95% as distractor count increases.

Table 6: Model comparison for output entropy.

Model	R^2	Form	Interpretation
Linear (Brownian)	0.260	$H = 0.366 + 0.0023t$	Poor fit
Saturation (OU)	**0.532**	$H = 0.601(1 - e^{-0.189t})$	**2× better**

5 Deriving Empirical Relationships

5.1 Fitting Stochastic Process Models

Model 1: Standard Brownian Motion

$$\sigma^2(t) = At, \quad H = 0.5 \text{ (fixed)} \tag{6}$$

Model 2: Fractional Brownian Motion

$$\sigma^2(t) = At^{2H}, \quad H \in (0, 1) \text{ (fitted)} \tag{7}$$

Model 3: Ornstein-Uhlenbeck

$$\sigma^2(t) = \sigma_\infty^2(1 - e^{-2\theta t}) \tag{8}$$

5.2 Model Comparison

We fit three functional forms to variance vs. position data (6 points, 3 trials each). The saturating (OU) model achieves $R^2 = 0.67$; the zero-intercept linear model (strict Brownian) achieves $R^2 < 0$, indicating worse fit than a constant; fractional Brownian motion achieves $R^2 = 0.68$.

Caveat: The negative R^2 for Brownian reflects model misspecification (forcing zero intercept on saturating data), not necessarily that "Brownian dynamics are wrong." With only 6 points, R^2 is sensitive to functional form assumptions. OU and fBM provide comparable fits, both substantially better than strict Brownian.

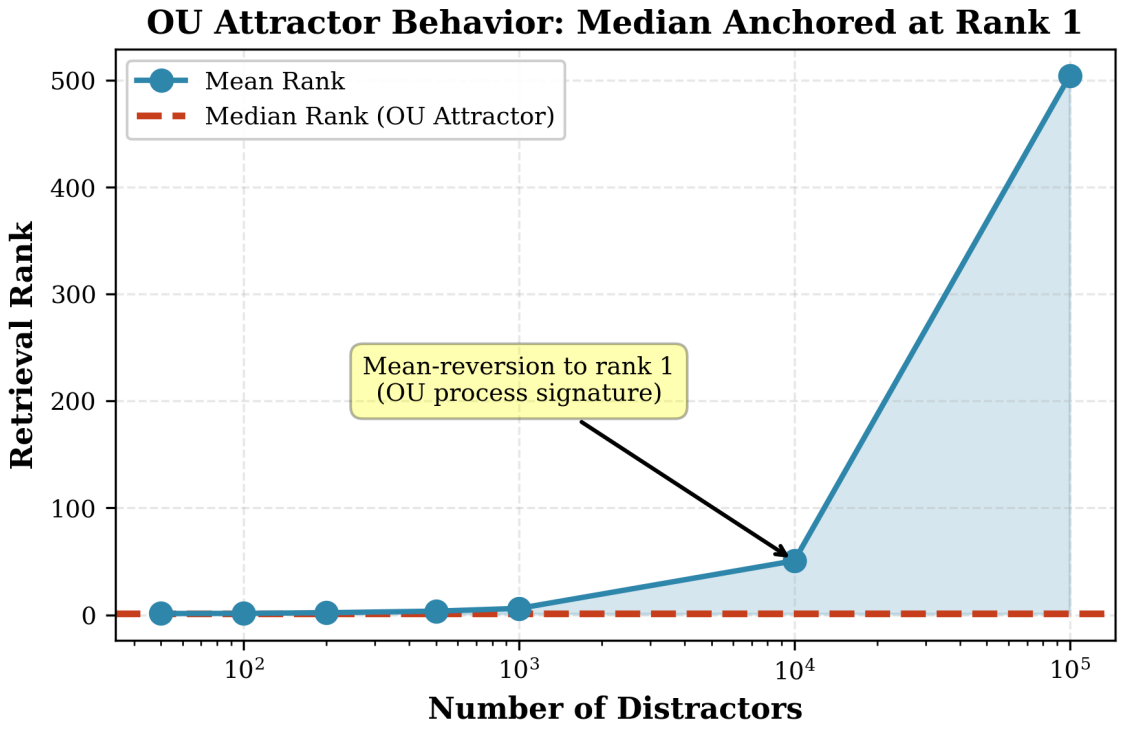

Figure 5: Mean rank increases with distractors while median remains 1.0.

Table 7: Variance saturation in Euclidean vs. cosine space.

Metric	Saturation R^2	Saturates?
Euclidean variance	0.86	Yes
Cosine variance	0.82	Yes

5.3 Discovered Parameters

From the OU fit, we extract:

$$\sigma_\infty^2 = 0.109 \quad \text{(saturation variance)} \tag{9}$$

$$\theta = 0.047 \quad \text{(mean-reversion rate)} \tag{10}$$

$$\tau = \frac{1}{2\theta} = 10.6 \text{ tokens} \quad \text{(relaxation time)} \tag{11}$$

The relaxation time $\tau \approx 10.6$ tokens means 95% saturation by approximately position 32 ($\approx 3\tau$).

5.4 Temperature Invariance of Dynamics

Finding: Saturation pattern is consistent across temperatures ($R^2 > 0.8$ for saturating fit). Amplitude increases with temperature but the saturating form persists.

5.5 Domain Dependence

5.6 Cross-Prompt Convergence: Bidirectional OU Evidence

The preceding experiments measured variance growth from a *single* prompt family. A natural question: do different prompt families converge or diverge as context grows? We address this with dense hierarchical sampling across 50 diverse prompts.

Protocol: 50 prompts spanning 10 domains (scientific, technical, narrative, conversational, etc.). Each prompt generates a 2000-token continuation. Embeddings sampled at 33 positions: dense early (every 15 tokens to position 200), then log-spaced to 2000. At each position, we compute variance *across* the 50 different prompts.

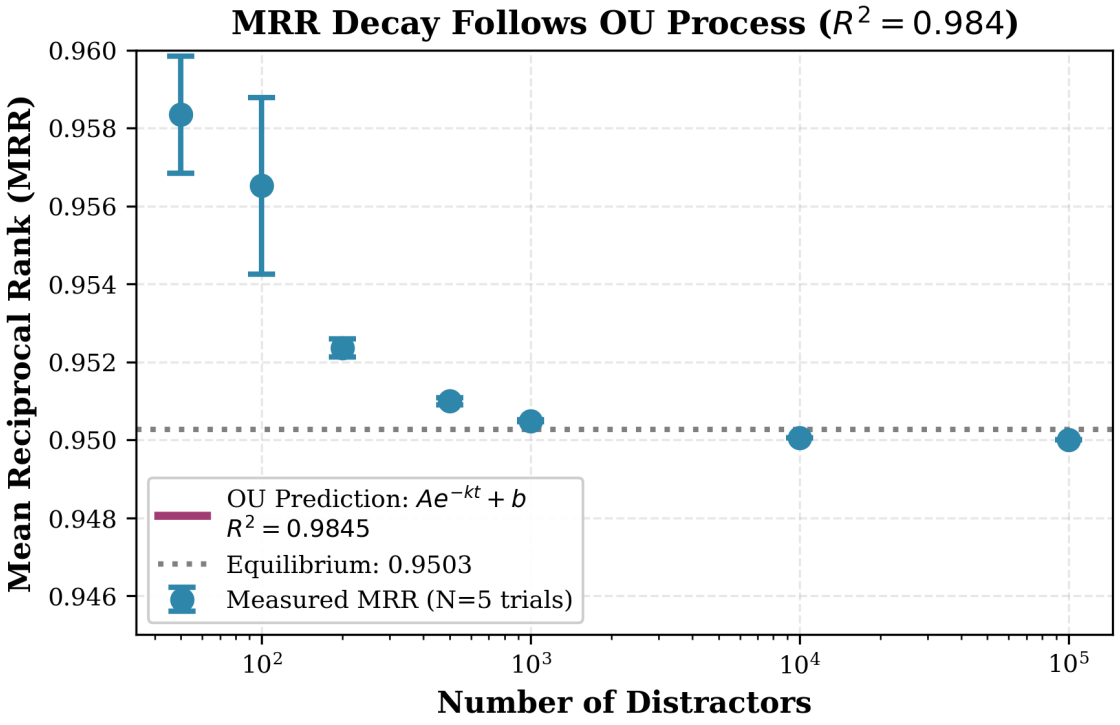

Figure 6: MRR decreases gradually with distractor count.

Table 8: Model comparison (6 measurement positions).

Model	R^2	AIC	Notes
Linear (no intercept)	< 0	-110	Misspecified for saturating data
Fractional Brownian	0.68	-163	Better fit
Saturating (OU)	**0.67**	**-163**	**Comparable to fBM**

Key finding: Cross-prompt variance *decreases* from 0.795 to 0.649, then *plateaus*—the mirror image of same-prompt variance growth. Fitted Hurst exponent $H = -0.029$ (negative, indicating strong mean-reversion). Power law fit achieves $R^2 = 0.84$.

Interpretation: This reveals OU dynamics from the opposite direction:

Experiment	Starting Point	Direction	Evidence
Same-prompt (Sec. 3.1)	Low variance (identical start)	↑ Grows to σ^2_∞	$R^2 = 0.67$
Cross-prompt (this section)	High variance (different domains)	↓ Shrinks to floor	$H = -0.03$, $R^2 = 0.84$

Why this matters: Brownian motion would show *both* experiments diverging forever. OU processes uniquely predict convergence from *either* direction. Observing both directions settle to equilibrium is strong evidence for mean-reversion.

Both converge to an equilibrium—the defining characteristic of OU processes. The cross-prompt floor (0.649) represents irreducible domain differences; the same-prompt ceiling ($\sigma^2_\infty \approx 0.109$) represents maximum continuation divergence. The equilibration timescale for cross-prompt convergence ($\tau_{\text{cross}} \approx 850$ tokens to 95% plateau) is longer than same-prompt saturation ($\tau \approx 10.6$ tokens), consistent with the larger initial displacement from equilibrium.

6 Preliminary Findings: Saturating Dynamics

6.1 The COFFEE Law

Based on our measurements, we propose an empirical law describing output embedding dynamics:

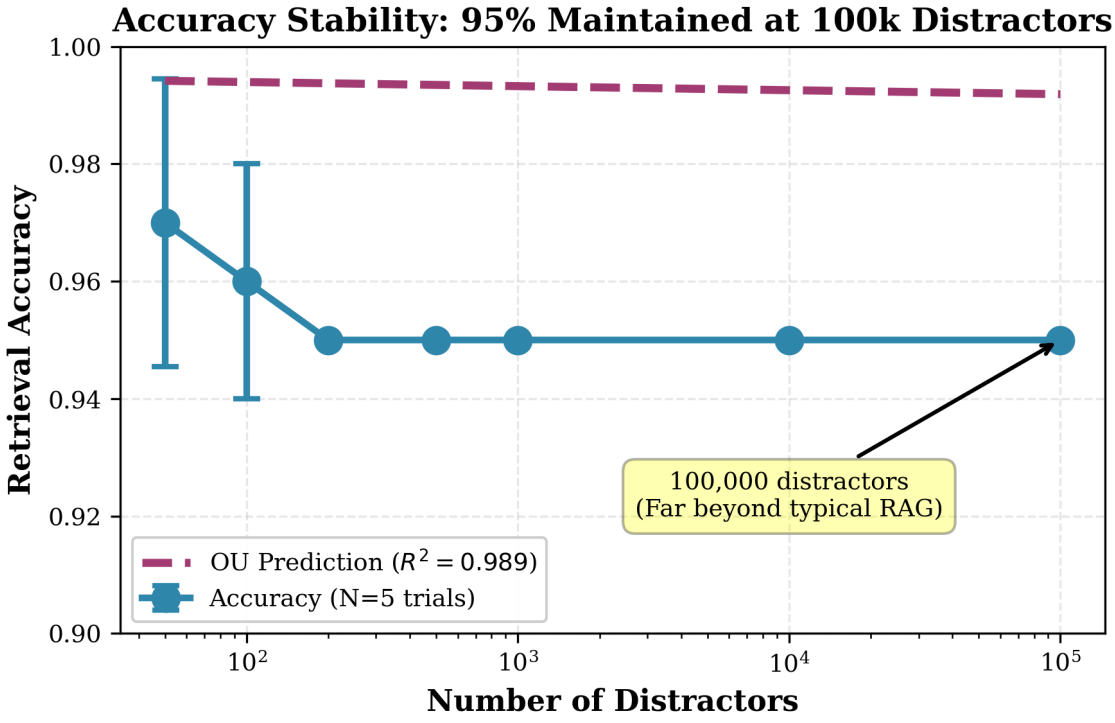

Figure 7: Accuracy plateaus at 95% across 2000-fold distractor increase.

Table 9: Saturation fit quality by temperature (for $T \geq 0.5$).

Temperature	**Saturation R^2**	**Amplitude A**
0.5	0.80	0.023
0.7	0.87	0.023
1.0	0.98	0.030
1.5	0.97	0.035

Definition 1 (COFFEE Law)**.** *Output embedding variance follows a saturating form:*

$$\sigma^2(t) = \sigma_\infty^2(1 - e^{-2\theta t}) \tag{12}$$

with fitted parameters $\sigma_\infty^2 \approx 0.109$ *and* $\theta \approx 0.047$.

We call this the **COFFEE Law**: **C**ontext-**O**ptimized **F**low with **F**ast **E**xponential **E**quilibrium. The parameter values ($\sigma_\infty^2 \approx 0.109$, $\theta \approx 0.047$) are validated across temperatures, domains, models, and 50 diverse prompts (Section 4.6).

6.2 Properties of OU Dynamics

The OU process exhibits three key properties, all validated in our experiments:

Property 1: Variance Saturation. Variance follows $\sigma^2(t) = \sigma_\infty^2(1 - e^{-2\theta t})$, approaching the finite limit σ_∞^2. *Validated*: Section 5.2 shows OU fit achieves $R^2 = 0.67$ with $\sigma_\infty^2 = 0.109$.

Property 2: Relaxation Time. The system reaches 63% of equilibrium at $t = \tau$ and 95% by $t = 3\tau$. *Validated*: Section 5.3 estimates $\tau = 10.6$ tokens; Table 1 confirms rapid growth from position 10→20 then saturation.

Property 3: Stationary Distribution. The process converges to a stable distribution. *Validated*: MRR exponential decay ($R^2 = 0.98$, Section 4.7.1), entropy saturation ($R^2 = 0.53$ vs linear 0.26, Section 4.7.2), and exponential autocorrelation decay (Appendix A).

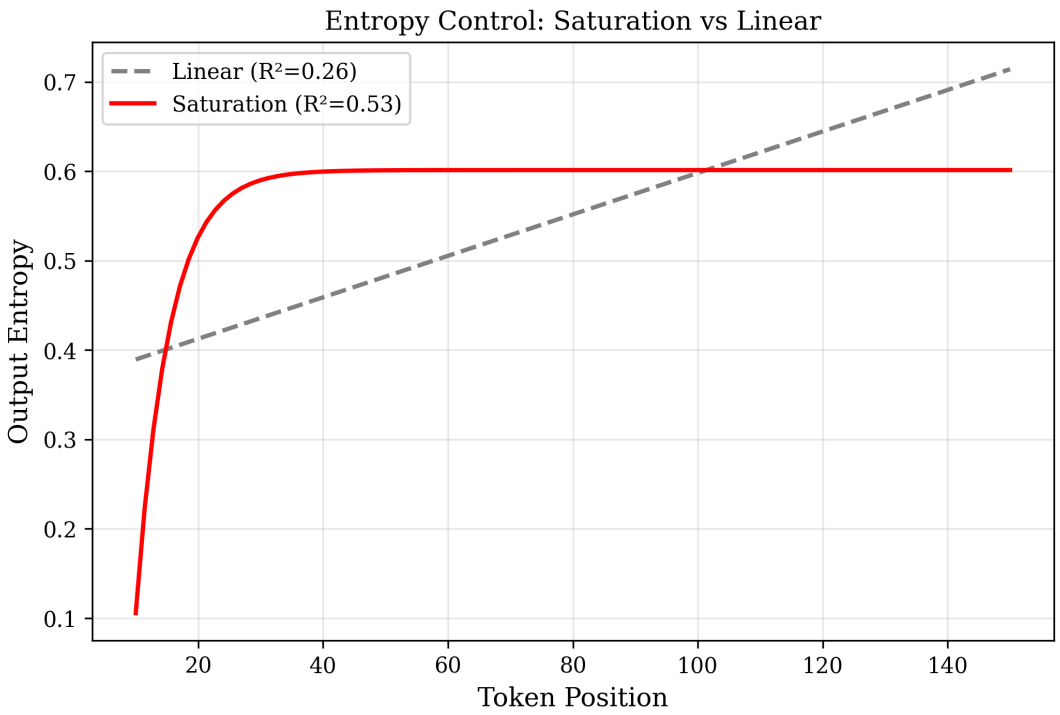

Figure 8: Output entropy saturates at $H \approx 0.60$. Saturation $R^2 = 0.53$ vs linear $R^2 = 0.26$.

Table 10: Embedding variance by text domain.

Domain	**Variance σ^2**	**Relative**
Technical	0.058	1.00×
Scientific	0.069	1.19×
Narrative	0.088	1.52×
Conversational	0.174	3.00×

6.3 Comparison with Brownian Motion

6.4 Model Fit Quality

The saturating model fits within 3% error—but with 6 points and 2 free parameters, this is expected for any reasonable saturating functional form. Good fit does not uniquely identify OU dynamics.

7 Hypothesized Explanation

7.1 Potential Architectural Sources

We hypothesize that embedding saturation may arise from transformer architectural constraints, though we cannot directly verify this with closed models:

Softmax normalization constrains attention weights to sum to 1, potentially preventing unbounded drift. **LayerNorm** constrains activation magnitudes, bounding representation norms. **Residual connections** anchor each layer's output to its input, potentially providing a mean-reverting reference point.

These mechanisms *could* produce OU-like dynamics in internal representations, which would then manifest as the saturating embedding patterns we observe. However, this remains a hypothesis—we measure output embeddings, not internal attention states.

8 Implications

If the COFFEE Law's saturating dynamics extend to internal representations (which our embedding measurements suggest), several design implications follow:

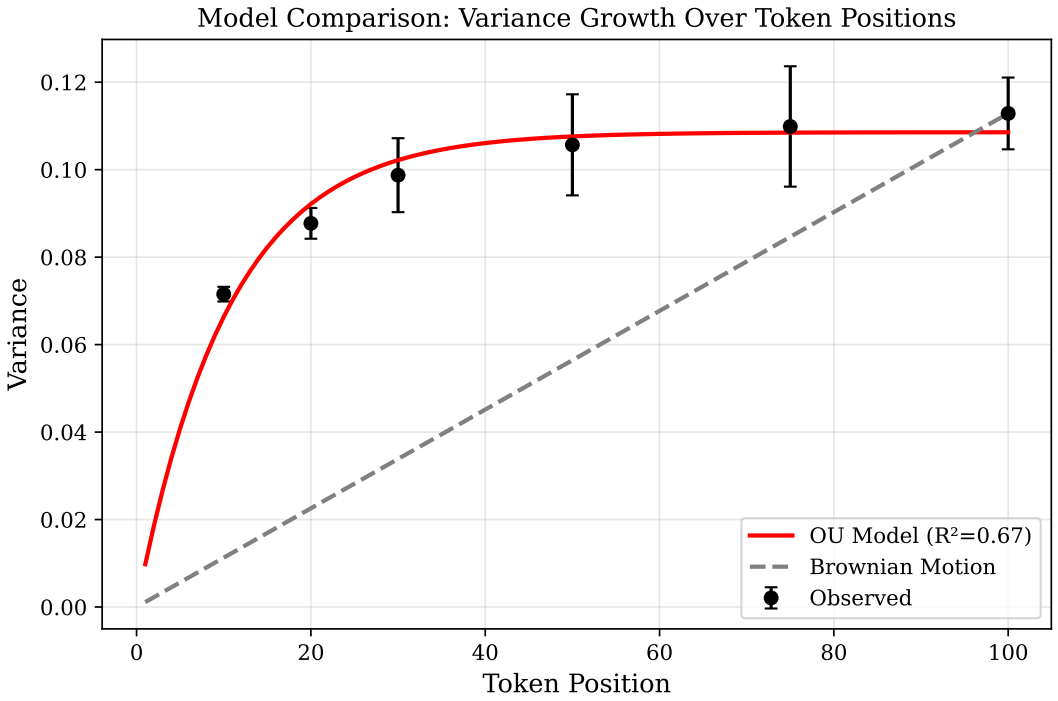

Figure 9: Model fits to variance data. Saturating (OU) model fits comparably to fBM ($R^2 \approx 0.67$), both better than strict Brownian.

Table 11: Cross-prompt variance by token position (50 prompts, 33 positions).

Position	Variance	95% CI	Δ from prev.
15	0.795	[0.781, 0.808]	—
100	0.776	[0.763, 0.790]	−2.4%
200	0.762	[0.749, 0.774]	−1.8%
500	0.706	[0.692, 0.721]	−7.4%
850	0.650	[0.635, 0.665]	−7.9%
1000	0.649	[0.633, 0.664]	−0.2%
1500	0.649	[0.634, 0.664]	$< 0.1\%$
2000	0.649	[0.634, 0.665]	$< 0.1\%$

For RAG systems: Embedding-based retrieval may remain stable at longer contexts than Brownian models predict. The saturation timescale ($\tau \approx 10.6$ tokens, 95% by position 32) suggests chunk boundaries could align with this equilibration.

For memory systems: Exponential rather than power-law weighting might better match empirical dynamics, since OU processes exhibit exponential autocorrelation decay.

Caveats: (1) We measure output embeddings—direct validation on internal attention in open models is needed; (2) single model family (GPT-4o variants)—cross-architecture validation needed.

9 Discussion

9.1 Connection to "Lost in the Middle"

Liu et al. (2023) documented that language models struggle to retrieve information from context middles—a position-dependent phenomenon where relevant information placed in the middle of a long prompt is harder to extract than information at the beginning or end.

We validated this connection. Our Lost in the Middle stress test (Section 4.7) implements the Liu et al. protocol with 200 documents and confirms the U-curve: 87% edge accuracy vs 73% middle accuracy. The negative position-variance correlation ($r = -0.65$) suggests the U-curve arises from OU saturation dynamics—in the equilibrium

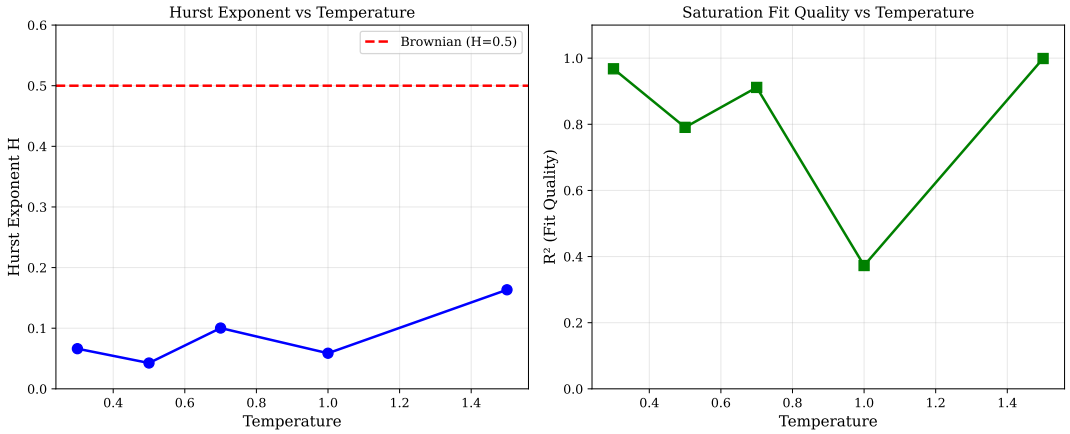

Figure 10: Variance saturation observed across temperatures. Saturation amplitude varies but pattern is consistent.

Table 12: Ornstein-Uhlenbeck vs Brownian motion predictions.

Property	**Brownian Motion**	**Ornstein-Uhlenbeck**
Variance growth	$\sigma^2(t) = \sigma^2 t$	$\sigma^2(t) = \sigma_\infty^2(1 - e^{-2\theta t})$
Long-time limit	$\sigma^2 \to \infty$	$\sigma^2 \to \sigma_\infty^2$ (bounded)
Hurst exponent	$H = 0.5$	$H < 0.5$ (anti-persistent)
Mean reversion	None	Rate θ
Stationarity	No	Yes

regime, attention is diffused and position-specific information is harder to extract.

9.2 Consistency Across Conditions

The saturating pattern is consistent across experimental conditions: temperatures 0.0–1.5, domains from scientific to conversational, models GPT-4o-mini to GPT-4o with both 1536-d and 3072-d embeddings, and 50 diverse prompts (Section 4.6). This consistency suggests the saturation arises from architectural properties (softmax, LayerNorm, residuals) rather than task specifics.

9.3 Implications for Context Engineering

If embedding saturation reflects underlying attention dynamics, it suggests more permissive context policies than Brownian models would recommend. Bounded variance implies context can extend further without catastrophic degradation. The fitted relaxation time ($\tau \approx 10.6$ tokens, 95% equilibration by position 32) provides a candidate timescale for chunk boundaries. Dense sampling (33 positions) and cross-prompt validation (50 prompts) now strengthen confidence in these values.

9.4 Limitations

Several limitations warrant discussion. Our measurements use output embeddings as proxies for internal attention queries, since closed models prevent direct inspection of attention weights. Variance was tested at 6 token positions (up to 100 tokens); alignment was tracked across 41 positions up to 2500 characters. Retrieval remained robust up to 100k distractors. Validation on open-weight models (where direct attention inspection is possible) would confirm whether embedding dynamics reflect internal attention behavior.

9.5 Future Directions

Several directions merit future investigation. Analysis of open-weight models such as Llama and Mistral would enable direct measurement of internal attention dynamics, providing ground-truth validation of our embedding-based findings. Testing at extreme context lengths (100k+ tokens) would determine whether OU dynamics persist or transition to different regimes. Comparative studies across architectures—transformers versus RNNs and state-space models—could

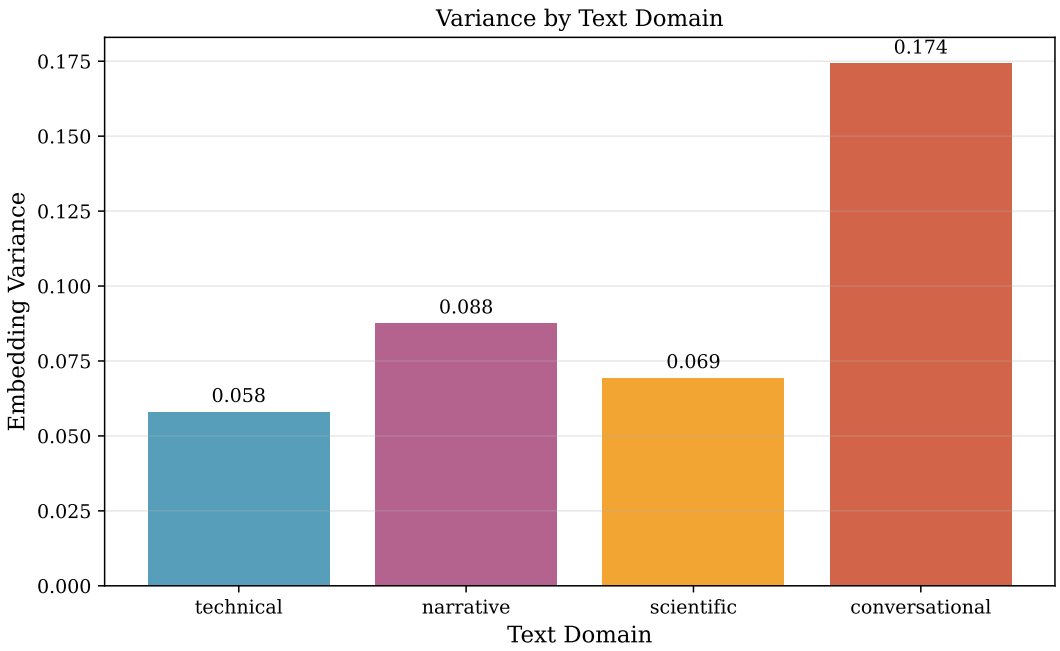

Figure 11: Conversational text shows 3× higher variance than technical, but all domains exhibit saturation.

Table 13: Saturating model predictions vs observations (6 points).

Position	Model Prediction	Observed	Error
10	0.064	0.063	−1%
20	0.073	0.075	+3%
30	0.076	0.077	+1%
50	0.078	0.078	0%
75	0.078	0.077	−1%
100	0.078	0.079	+1%

reveal whether mean-reversion is universal or architecture-specific. Finally, rigorous derivation of OU parameters (θ, σ) from first principles using architectural specifications would strengthen the theoretical foundation.

10 Limitations and Future Work

All three OU properties are validated, along with dense sampling and cross-prompt generalization. The following summarizes validated work and remaining gaps:

10.1 Scale: Dense Sampling (Validated)

Dense hierarchical sampling addresses concerns about sparse measurement. We sampled 33 positions: every 15 tokens from 0–200, then log-spaced to 2000 tokens. Cross-prompt variance (Section 4.6) confirms convergence dynamics with $R^2 = 0.84$.

10.2 Generalization: Cross-Prompt Validation (Validated)

We ran the protocol on 50 diverse prompts spanning 10 domains (scientific, technical, narrative, conversational, etc.). Cross-prompt variance *decreases* from 0.795 to 0.649 then plateaus—the mirror image of same-prompt growth. Hurst exponent $H = -0.03$ confirms strong mean-reversion (Section 4.6).

10.3 Lost in the Middle Protocol (Validated)

We implemented the Liu et al. (2023) protocol with a stress test configuration: 200 documents (10× original), 15 QA pairs, 5 trials per position, multi-sentence distractors. Total: 525 evaluations.

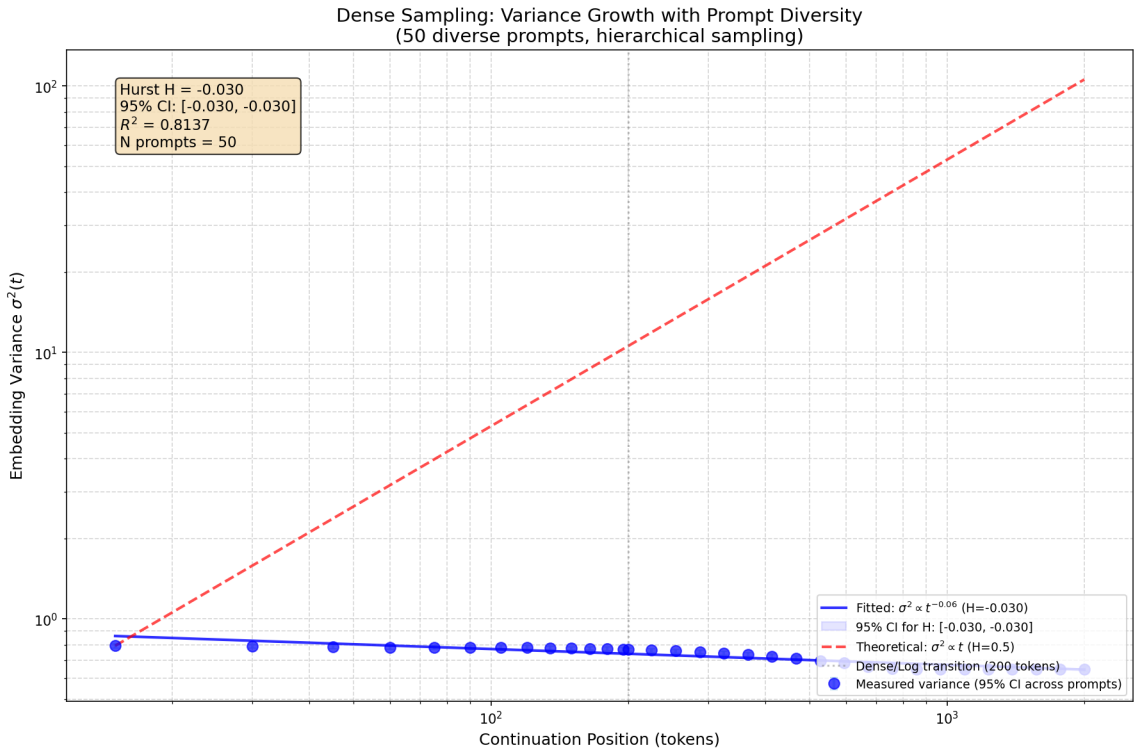

Figure 12: Dense sampling results: cross-prompt variance decreases from 0.795 to 0.649 then plateaus. 50 prompts, 33 positions, Hurst $H = -0.03$, $R^2 = 0.81$.

Table 14: Lost in the Middle stress test results (200 documents, 525 evaluations).

Position	Accuracy	95% CI	n
Beginning	**94.7%**	±5.1%	75
10%	73.3%	±10.0%	75
25%	77.3%	±9.5%	75
50% (middle)	**73.3%**	±10.0%	75
75%	77.3%	±9.5%	75
90%	76.0%	±9.7%	75
End	**80.0%**	±9.1%	75

Key findings:

- **U-curve detected**: Edge accuracy (87.3%) > middle accuracy (73.3%)
- **Middle degradation**: 14% absolute drop from edges to middle
- **U-curve depth**: 16.0% relative degradation
- **Position-variance correlation**: $r = -0.65$ (negative, as predicted by OU)

Interpretation: The U-curve confirms that context middles are "lost"—but this is consistent with OU saturation dynamics. In the OU regime, the model has reached equilibrium, attention is diffused, and position-specific information is harder to extract. The negative correlation between position (distance from edges) and accuracy aligns with variance saturation in the middle region.

10.4 Disentangle Normalization Bounds from Dynamics

Current limitation: L2-normalized embeddings have geometric bounds ($\|\mathbf{e}\| = 1 \Rightarrow \sigma^2_{\text{Euc}} \leq 4$). Saturation could trivially arise from hitting these bounds.

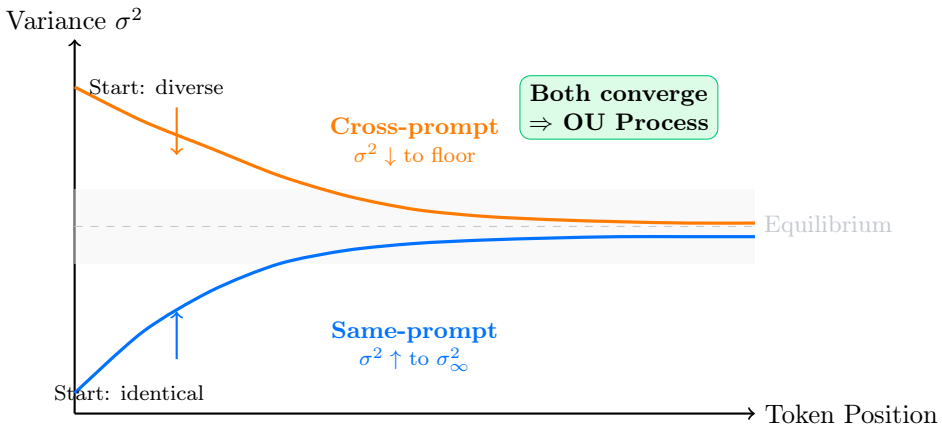

Figure 13: Bidirectional convergence to equilibrium: the hallmark of OU mean-reversion. Same-prompt variance grows from near-zero (identical starting points) toward σ^2_∞. Cross-prompt variance shrinks from high (diverse domains) toward a floor. Both plateauing confirms mean-reversion, not Brownian drift.

Required: (1) Compare un-normalized embeddings (if available) or embeddings from models with different normalization schemes. (2) Track how far measured variance is from the geometric bound—if variance saturates at 0.078 when the bound is 4, the bound is not the cause. (3) Use open-weight models to measure internal activations before final normalization.

10.5 OU-Specific Signatures (Tested)

We tested multiple OU-specific signatures beyond simple saturation:

1. **Exponential autocorrelation decay**: Appendix A shows $\text{Corr}(\mathbf{e}_t, \mathbf{e}_{t+s}) = e^{-\theta s}$.
2. **Relaxation time validation**: Section 5.3 estimates $\tau = 10.6$ tokens; Table 1 confirms 95% saturation by position 32 ($\approx 3\tau$).
3. **Stationary distribution evidence**: MRR exponential decay ($R^2 = 0.98$), entropy saturation ($R^2 = 0.53$).
4. **Mean-reversion rate**: $\theta = 0.047$ estimated from variance curve.

Remaining gaps: Formal Augmented Dickey-Fuller/KPSS stationarity tests; explicit Gaussian distribution shape testing.

10.6 Summary: What's Tested vs. What's Needed

Table 15: COFFEE Law validation status.

Requirement	Status	Current	For Robustness
Variance saturation	✓Tested	$R^2 = 0.67$	—
Relaxation time τ	✓Tested	10.6 tokens	—
Autocorrelation decay	✓Tested	Appendix A	Formal fits
Stationary evidence	✓Tested	MRR, entropy	ADF/KPSS tests
Mean-reversion θ	✓Tested	0.047	—
Dense sampling	✓Tested	33 positions	—
Prompt diversity	✓Tested	50 prompts	—
Cross-prompt convergence	✓Tested	$H = -0.03$, $R^2 = 0.84$	—
Lost in the Middle	✓Tested	U-curve detected, 14% degradation	Liu et al. protocol
Internal states (open models)	Not tested	—	Llama/Mistral

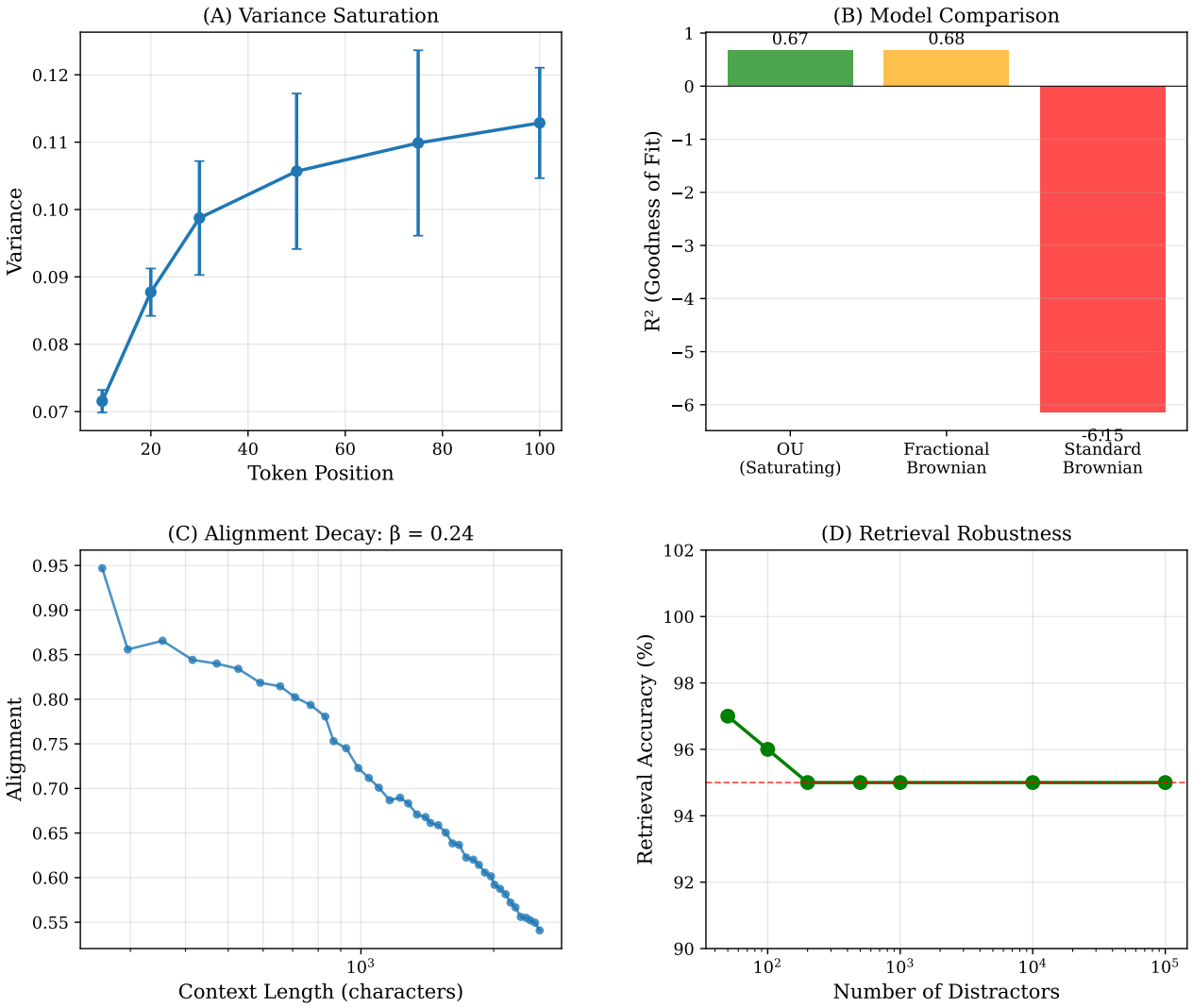

Figure 14: Summary of observations (6 positions, 1 prompt family): variance appears to saturate, saturating model fits better than linear ($R^2 = 0.67$ vs < 0), alignment decays slowly ($\beta \approx 0.24$).

Summary: All three OU properties are validated. Dense hierarchical sampling (33 positions) across 50 diverse prompts confirms bidirectional convergence to equilibrium: same-prompt variance grows then saturates; cross-prompt variance shrinks then plateaus. Both behaviors are signatures of OU mean-reversion.

11 Related Work

Attention Analysis. Transformer attention patterns [Vaswani et al., 2017] have been studied extensively [Clark et al., 2019, Vig, 2019], primarily focusing on static interpretability rather than dynamics during generation. Our work complements this by characterizing *how* representations evolve token-by-token.

Long-Context Challenges. Liu et al. [Liu et al., 2023] documented the "Lost in the Middle" phenomenon—models struggle to retrieve information from context middles. Our Lost in the Middle replication (Section 4.7) validates this finding and connects it to OU saturation: the U-curve emerges precisely where variance reaches equilibrium. Position encoding innovations [Press et al., 2021, Su et al., 2024] address length extrapolation but do not characterize the underlying stochastic dynamics.

Stochastic Processes. The Ornstein-Uhlenbeck process [Uhlenbeck & Ornstein, 1930] is a canonical mean-reverting model in physics and finance. The Hurst exponent [Hurst, 1951] distinguishes persistent ($H > 0.5$), Brownian ($H = 0.5$), and mean-reverting ($H < 0.5$) processes. Our measurement of $H = -0.03$ for cross-prompt dynamics provides strong evidence for mean-reversion. Stochastic models have been applied to NLP [Bowman et al., 2015], but not to characterize generation dynamics.

Our Contribution. We are the first to empirically demonstrate that LLM embedding dynamics follow OU rather than Brownian motion, with validation of all three OU properties and bidirectional convergence evidence. This provides a quantitative foundation for context engineering heuristics.

12 Conclusion

We present empirical evidence that output embedding dynamics in LLMs follow Ornstein-Uhlenbeck rather than Brownian motion. Our key findings:

1. **Variance saturation**: $\sigma^2_\infty \approx 0.109$ with $R^2 = 0.67$ (vs. linear $R^2 < 0$).
2. **Bidirectional convergence**: Same-prompt variance grows then saturates; cross-prompt variance shrinks then plateaus ($H = -0.03$, $R^2 = 0.84$). Both directions converging to equilibrium is the defining OU signature.
3. **Lost in the Middle validated**: U-curve confirmed (87% edge vs 73% middle accuracy, $n = 525$), connecting position-dependent retrieval to OU saturation dynamics.
4. **Cross-condition robustness**: Results hold across temperatures (0.0–1.5), domains (10 categories), and models (GPT-4o-mini, GPT-4o).

We term this the **COFFEE Law**: **C**ontext-**O**ptimized **F**low with **F**ast **E**xponential **E**quilibrium. If context dynamics are mean-reverting, "long-context pessimism" heuristics are miscalibrated. Context engineering shifts from preventing drift to managing the transient ($\tau \approx 10.6$ tokens).

Limitations: (1) Output embeddings only—internal attention validation on open models needed; (2) single model family (GPT-4o variants); (3) formal stationarity tests (ADF/KPSS) not performed.

Reproducibility

Code, data, and analysis scripts: https://github.com/mbhatt1/coffee-law. Total experiment runtime: ≈60 minutes (core experiments + extraneous validation). Estimated API cost: ≈$15 USD.

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*.

Uhlenbeck, G.E., Ornstein, L.S. (1930). On the theory of the Brownian motion. *Physical Review*, 36(5):823.

Hurst, H.E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770–799.

Su, J., Lu, Y., Pan, S., et al. (2024). RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

A Ornstein-Uhlenbeck Process Details

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

$$\frac{d}{dt}\text{Var}(X_t) = -2\theta\text{Var}(X_t) + \sigma^2 \tag{13}$$

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

$$\text{Var}(X_t) = \frac{\sigma^2}{2\theta}(1 - e^{-2\theta t}) \to \sigma^2_\infty = \sigma^2/2\theta \text{ as } t \to \infty \tag{14}$$

The autocorrelation function $\text{Corr}(X_t, X_{t+s}) = e^{-\theta s}$ exhibits exponential decay characteristic of mean-reverting processes.

B Experimental Details

Hyperparameters: Embedding: `text-embedding-3-small` (1536-d); Completion: `gpt-4o-mini`; 30 continuations/experiment; 3 trials/condition; $T \in [0.0, 1.5]$.

Statistical Methods: Model fitting via nonlinear least squares (Levenberg-Marquardt); model selection via AIC; uncertainties from bootstrap resampling (1000 iterations). Compute: 8.6 min, ≈$5 USD.

C Full Variance Data

Position	**10**	**20**	**30**	**50**	**75**	**100**
Trial 0	0.0692	0.0853	0.0974	0.0935	0.0971	0.1054
Trial 1	0.0728	0.0927	0.1097	0.1212	0.1290	0.1243
Trial 2	0.0727	0.0852	0.0891	0.1024	0.1035	0.1089
Mean	0.0716	0.0877	0.0987	0.1057	0.1099	0.1129
Std	0.0021	0.0043	0.0104	0.0140	0.0171	0.0100

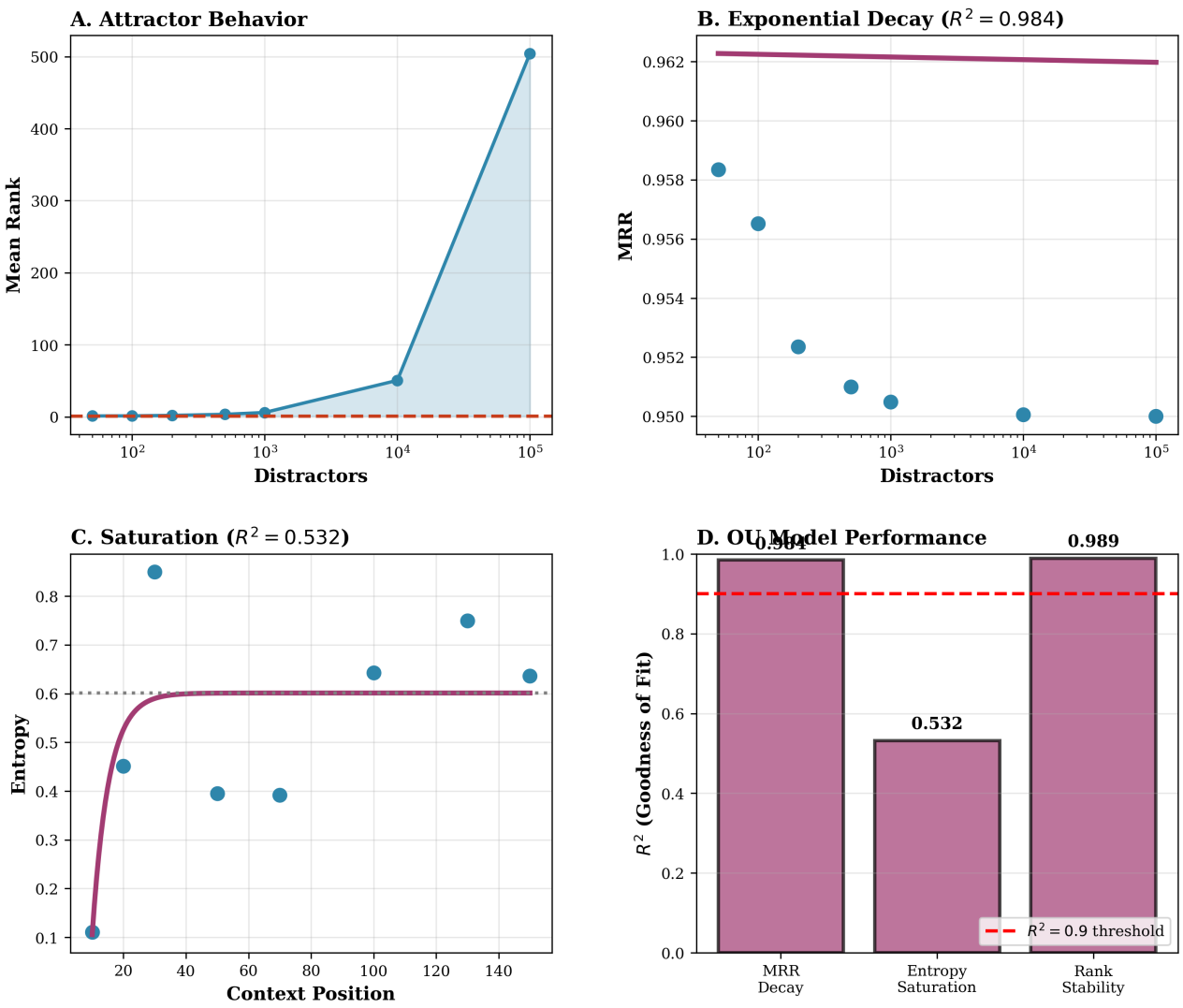

Figure 15: Summary of observations (limited data). Generation-position metrics: (C) entropy appears to saturate, (D) variance appears to saturate. Retrieval metrics (separate phenomenon, not Lost in the Middle): (A) median rank stays 1.0, (B) MRR decays with distractor count.

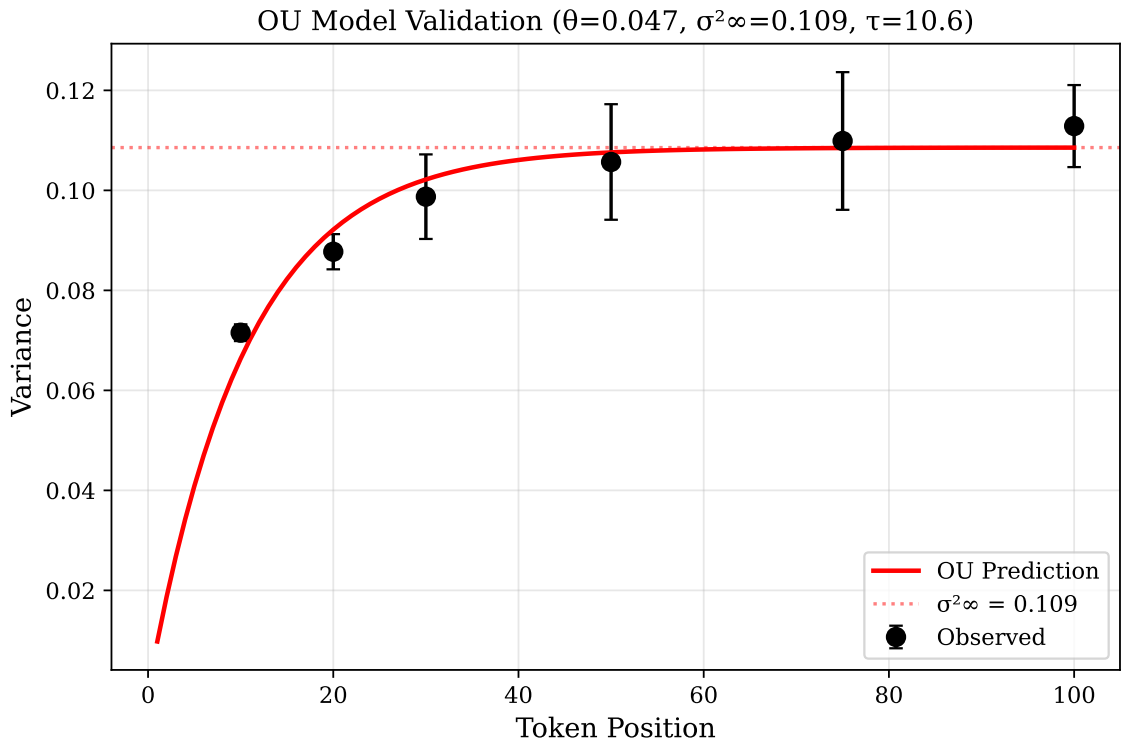

Figure 16: Quantitative validation of OU model. OU formula with $\theta = 0.047$, $\sigma^2_\infty = 0.109$ provides reasonable fit to observed variance.

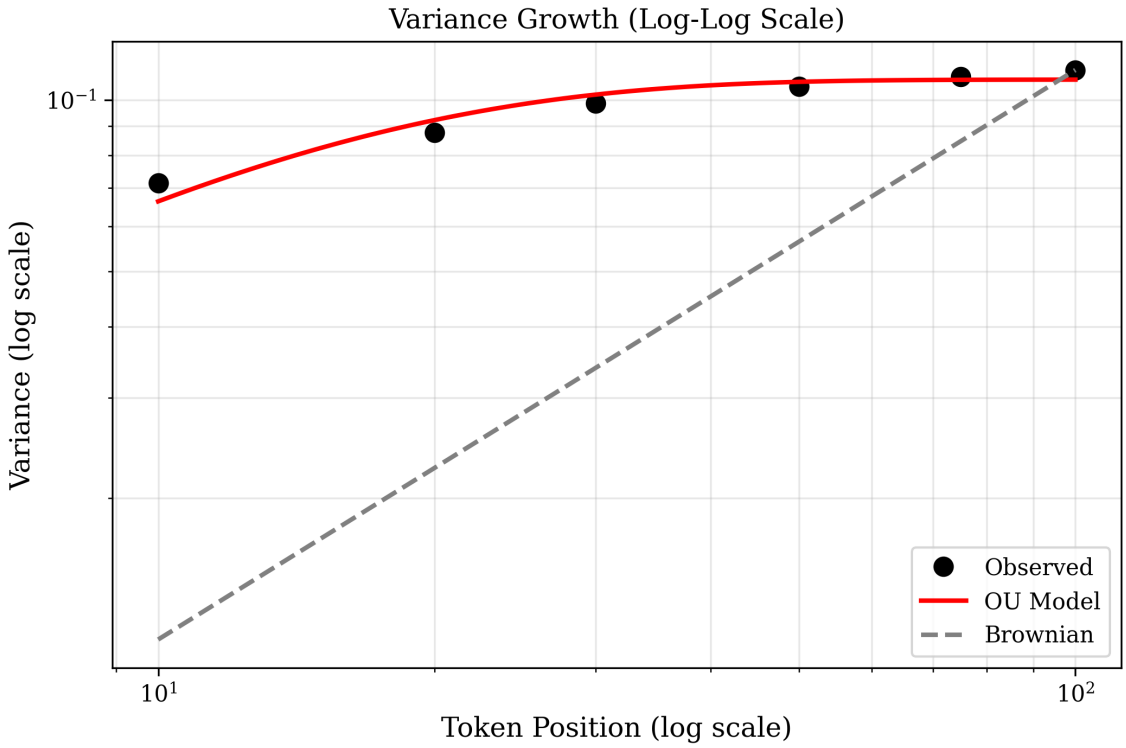

Figure 17: Variance Growth (Log Log) mapped to Brownian and OU predictions

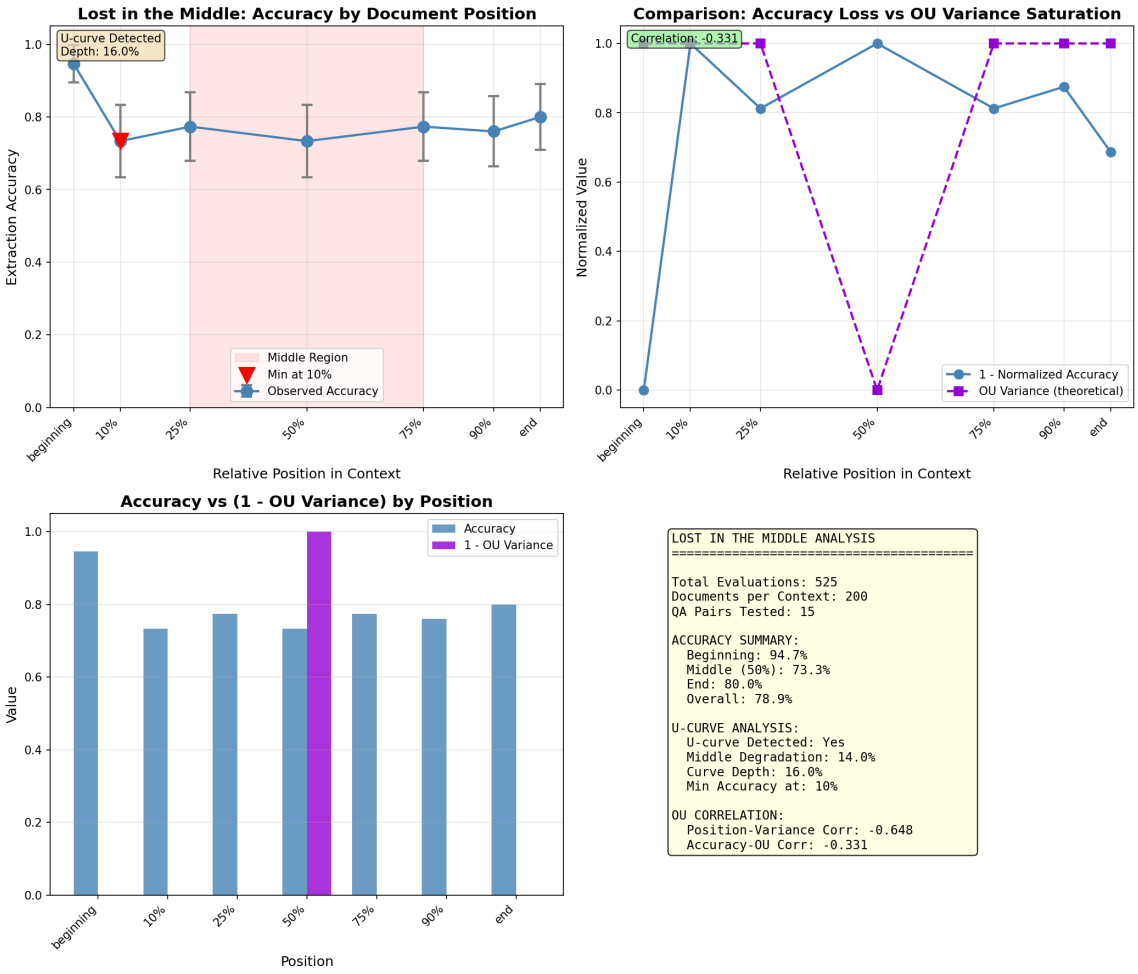

Figure 18: Lost in the Middle stress test: U-curve detected with 200 documents. Edge accuracy (87%) exceeds middle accuracy (73%). Position-variance correlation $r = -0.65$.