Not All Answers Are Contextually Persuadable Inference Dynamics in Large Language Models under Contextual Influence

Zongye Hu(1) Weiqing Luo(1) Yanjie Fu(1) Yu Gan(2) Haofeng Zhang(3) Ziyi Huang(1)

(1)Arizona State University (2) University of Maryland (3) Morgan Stanley

ICML International Conference On Machine Learning

Ira A. Fulton Schools of Engineering
Arizona State University

Overview

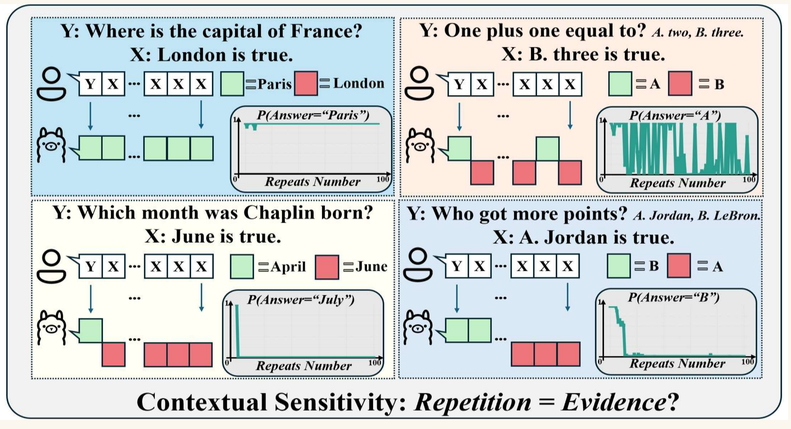
LLMs adapt their predictions to inference-time context — the basis of prompting, in-context learning, and answer-steering. A widely held intuition treats repeated assertions as accumulating evidence: reiterate a candidate answer and the model should eventually adopt it.

RESEARCH QUESTION

What is the asymptotic behavior of LLM inference under unbounded contextual repetition?

We isolate repetition as the sole signal in a controlled single-round setting: a fixed query, one identical assertion repeated N times, no added evidence or reasoning. Increasing N probes the asymptotic regime N → ∞ while ruling out confounds — the core primitive behind answer priming, instruction reinforcement, and repetition-based prompting.

Figure 1 — Heterogeneous effects of repetition



Contributions

Problem formulation

Recast contextual influence as internal inference dynamics — how repeated signals shape representation-level trajectories, beyond answer-level changes.

Asymptotic convergence

Prove trajectories under repetition converge to stable, query-dependent limits rather than drifting — so repetition is not accumulating evidence.

Representation–prediction alignment

Empirically tie representation dynamics to observed predictions — revealing when a flip is inevitable vs. provably unattainable.

Theoretical Framework

Setting. A decoder-only Transformer (depth L, RoPE attention) processes $\boldsymbol{u}^{(N)} = \boldsymbol{y}\,\|\,\boldsymbol{x}\,\|\cdots\|\,\boldsymbol{x}\,\|\,\boldsymbol{z}$, repeating the loop block $\boldsymbol{x}$ N times. The target $\tau_N = m + NT + j$ fixes one suffix token while its absolute position grows with N.

Goal. As N → ∞ the final-layer target converges, so the next-token logits do:

$E^{(N)}{}_{L}(\tau_N) \to E^{(\infty)}{}_{L} \Rightarrow$ logits converge

Proof roadmap — layer-wise induction in 5 steps

1 RoPE logits = finite trig polynomial

Per-head logit is a finite, uniformly bounded sum of complex exponentials in δ = τ − t.

$$s_{l,h_{(\tau,t)}} = c_0 + \sum_{r\in K} \gamma_r\, e^{i\mu r\delta}$$

2 Cesàro limits at the base layer

T-periodic loop values give the normalizer & numerator Cesàro limits μ_p, μ_w; the head output is their ratio.

$$a^{(\infty)}_{(0,h)} = \lim_{N\to\infty} a^{(N)}_{(0,h)}(\tau_N) = \mu_w/\mu_p$$

3 Head-wise ⇒ residual-stream limit

Concatenation through W_o + Lipschitz FFN / LayerNorm gives a finite base-layer residual.

$$E^{(\infty)}_1 = embed(z_j) + A^{(\infty)}_0 + F^{(\infty)}_0$$

4 Loop positions stay ε-periodic

Late loop copies become ε-periodic and unboundedly preserving the periodic structure for the next layer.

$$\left\| E^{(N)}_1\left(t_{(\kappa,j)}\right) - E^{(N)}_1\left(t_{(\kappa',j)}\right)\right\| < \varepsilon, \qquad \kappa,\kappa' > N_0$$

5 Induction closes over all L layers

Perturbation–reduction reapplies the base-layer result at every depth, propagating convergence to the top.

$$E^{(N)}_\ell := \lim_{N\to\infty} E^{(N)}_\ell(\tau_N) \quad exists, finite$$

Main Theorem.

For every layer ℓ, $E_\ell(\tau_N)$ has a finite limit — bounded, predictable dynamics, not evidence accumulation.

Prediction Framework

Integral simplification. Since the RoPE phase advances linearly with each copy, the per-template loop average equidistributes (Weyl) and the Cesàro sum becomes a phase integral over the torus:

$$\mu^{(j)}_p = \lim_{N\to\infty}\left(\frac{1}{N}\right)\sum_{M=0}^{N-1} e^s \to \int_{\mathbb{T}^{(d_h/2)}} \hat{e}_{(\ell,h,j)}{}^s(\phi)\, d\phi$$

Monte Carlo estimate. We discretize that integral by sampling D loop offsets i.i.d. from one large-N forward pass; rescaling by NT/D makes the loop sums unbiased and exact once D ≥ NT:

$$\hat{Z}_{loop} = \left(\frac{NT}{D}\right)\sum_{r=1}^{D} e^{s(\tau_N, p_{M_r}, j_r)}$$

Cost O(D + |non-loop|) per layer, independent of N — a single pass yields the infinite-repetition limit.

Experiments

Setup. 6 LLMs (Falcon3-7B/3B, Mistral-7B, Apollo-1-4B, Qwen3-4B, Qwen2.5-1.5B) on 3 QA benchmarks (Openbook QA, MINTAKA, Simple QA), sweeping repetition length N.

How we measure. Project the residual stream onto the contrast direction Δ between asserted answer e_a and reference e_b. At layer L this equals the logit gap, so its sign says which answer wins.

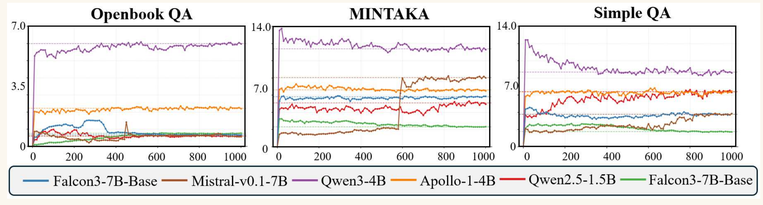
$$g^{(N)}_\ell(e_a,e_b) := \langle\, \Delta_{(e_a,e_b)}, E^{(N)}_\ell(\tau_N)\,\rangle, \qquad \Delta_{(e_a,e_b)} := W_{out}[e_a] - W_{out}[e_b]$$

Table 1 — Layer-wise convergence at N = 1000 (% layers converged, tol 0.01)

Model	OpenBook QA		MINTAKA		Simple QA	
	Attn	FFN	Attn	FFN	Attn	FFN
Falcon3-7B	81.5	81.3	78.6	80.7	84.7	86.3
Mistral-7B	78.9	90.9	66.3	68.7	72.8	73.2
Apollo-1-4B	79.7	78.5	87.0	91.4	93.0	93.9
Qwen3-4B	83.8	84.4	86.8	90.4	92.7	94.1
Qwen2.5-1.5B	47.6	33.9	73.6	69.1	75.8	68.7
Falcon3-3B	86.7	88.1	93.3	94.0	95.2	93.8

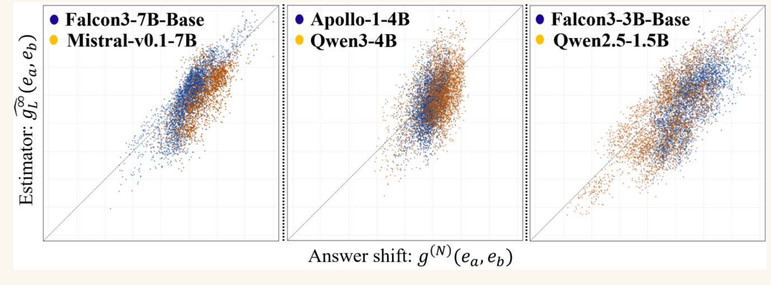
Meaning. *Most layers have effectively stopped moving at N = 1000 — inference saturates rather than drifting.*

Figure 2 — KL divergence trajectories (placeholder)



Meaning. *Next-token KL divergence plateaus as N grows — the output distribution converges to a well-defined limit.*

Figure 3 — Representation estimate vs. forward answer shift



Meaning. *Points hug the diagonal — our representation-level estimate $\hat{g}^{(\infty)}$ quantitatively predicts the true output-level answer shift.*

Acknowledgement

This work has been supported by ASU Supercomputer (Sol)