NOT ALL WHO WANDER ARE LOST: HALLUCINATIONS AS NEUTRAL DYNAMICS IN RESIDUAL TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

We separate onset from persistence and prove that persistence follows from the neutral dynamics of pre-LayerNorm residual transformers. Exact operator norms for LayerNorm, residual blocks, and the softmax decoder yield conservative upper bounds showing the absence of contractive or expansive bias at the decoded level. These bounds are sharpened by working with corridor constants that remain explicit and falsifiable. For open probes, drift decomposes into a predictable component bounded by the sharpened corridor and a centered martingale component controlled by concentration and central limit arguments. Neutrality is then lifted from paired rollouts to populations by casting trajectories or blocks as exchangeable agents in a mean-field game, yielding a population-invariant stable under depth and width scaling. Predictions are tested with controlled randomization audits up to `GPT2-large`: closed probes are centered and behave as bounded martingale differences, while open probe drift stays within the predicted corridor with magnitudes consistent with the sharper constants. Together, these theoretical and empirical results provide the first structural account of persistence, explaining why hallucinations persist across model scales without re-auditing hundreds of millions of parameters, and showing that interventions, which do not alter the residual backbone, cannot eliminate it once onset has occurred.

1 INTRODUCTION

Hallucinations remain one of the most persistent challenges for large language models. They arise in two phases: onset, when entropy or misspecification produces the first divergence between trajectories, and persistence, when that divergence continues to propagate step after step through autoregressive decoding. Recent advances have tackled onset through training and inference interventions such as scheduled sampling (Bengio et al., 2015; Mihaylova & Martins, 2019), sequence-level objectives (Ranzato et al., 2015), retrieval augmentation (Lewis et al., 2020), reinforcement learning with human feedback (Ouyang et al., 2022), and tool use (Schick et al., 2023). These methods improve factual accuracy and reduce exposure bias, but the deeper structural question remains open: once a deviation has occurred, why does it persist? Since training- and inference-time fixes do not alter residual dynamics, neutrality means onset errors inevitably persist, limiting what such strategies can achieve.

Existing diagnostics shed light on symptoms but not laws. Surface inconsistencies and semantic entropy provide useful signals (Farquhar et al., 2024; Lin et al., 2021; Manakul et al., 2023; Chen et al., 2024; Mündler & colleagues, 2024), and surveys emphasize the importance of human evaluation for factuality (Maynez et al., 2020; Kryściński et al., 2020; Ji et al., 2023; Huang et al., 2025). Yet these approaches remain empirical: they indicate when hallucinations happen, but they do not explain what the residual transformer backbone predicts about their evolution. The distinction between diagnostic cues and structural invariants is crucial. In this paper we ask: what dynamical law governs divergence inside autoregressive transformers?

OUR CONTRIBUTION AND NOVELTY

We show that persistence is the consequence of *neutral dynamics* in pre–LayerNorm residual transformers. Our approach combines exact architectural analysis, statistical inference, and a mean-field lift:

- We prove that the residual stack is neutral: paired rollouts exhibit neither contractive pull nor expansive push in expectation. This neutrality is certified by explicit Lipschitz constants for LayerNorm, the residual kernel, and the softmax decoder.
- We establish a blended inference rule: drift decomposes into a predictable corridor determined by architecture and a centered martingale part controlled by concentration inequalities. This makes neutrality a falsifiable claim with finite-sample guarantees.
- We lift neutrality from local probes to the population level by recasting trajectories or residual blocks as agents in a mean-field game. Under exchangeability, neutrality propagates to the limit, showing that persistence is scale-stable without re-auditing billions of parameters.

This integration of operator analysis, martingale inference, and mean-field scaling is, to our knowledge, the first structural account of hallucination persistence. It reframes the problem from an empirical irregularity into an architectural invariant, providing a principled explanation for why hallucinations endure across model scales.

Relation to prior work. Our perspective complements three active research fronts. First, detection and mitigation methods address onset by exploiting inconsistencies, semantic entropy, or external anchoring (Ranzato et al., 2015; Lewis et al., 2020; Ouyang et al., 2022; Schick et al., 2023; Farquhar et al., 2024). Second, stability analyses of residual networks and normalization layers provide the operator-theoretic tools we rely on (He et al., 2016; Haber & Ruthotto, 2017; Chen et al., 2018; Xiong et al., 2020; Miyato et al., 2018; Gouk et al., 2021). Third, mean-field methods from economics and probability (Lasry & Lions, 2007; Huang et al., 2006; Sznitman, 1991; Carmona & Delarue, 2018) have recently been adapted to neural networks and transformers (Yang et al., 2018; Fabian et al., 2024; **?**; Tembine et al., 2024). Our novelty lies in combining these strands into a unified framework that turns architectural constants into predictive laws, links them to falsifiable probe designs, and scales them through a mean-field lift.

2 BACKGROUND

To analyze persistence we formalize the autoregressive setting of pre–LayerNorm transformers and define how divergence between paired rollouts is measured, where a *rollout* is a sequence of hidden states and decoded tokens generated under autoregression.

Autoregressive dynamics. Large language models generate sequences by autoregression. Hidden states are $h_t \in \mathbb{R}^d$, and decoded token distributions are p_t living on the probability simplex

$$\Delta^{V-1} \;=\; \{p \in \mathbb{R}^V :\; p_i \geq 0,\; \sum_{i=1}^{V} p_i = 1\}.$$

A pre–LayerNorm (LN) residual stack applies blocks of the form

$$H_\ell(x) = x + G_\ell(\mathrm{LN}(x)), \qquad F = H_L \circ \cdots \circ H_1,$$

where $\mathrm{LN}(x) = \gamma \odot \frac{x-\mu(x)\mathbf{1}}{\sigma(x)} + \beta$ denotes LayerNorm (normalize to zero mean and unit variance, then rescale by γ and shift by β) (Ba et al., 2016). The decoder is affine plus a temperature–T softmax,

$$S(h) \;=\; \mathrm{softmax}_T(Wh + b), \qquad \mathrm{softmax}_T(z)_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}.$$

Together, (F, S) define a Markov autoregression

$$h_{t+1} = F(h_t), \qquad p_t = S(h_t), \qquad \tau_t \sim p_t,$$

which is the state space setting for our analysis in Section 3. For paired rollouts we maintain a second arm $(\tilde{h}_t, q_t)$ evolving under the same kernel, with $q_t = S(\tilde{h}_t)$ and $\tilde{\tau}_t \sim q_t$. Formal derivations are given in Appendices A.1–A.1.

Divergence and drift. To track how paired rollouts separate, we measure decoded divergence using the Jensen–Shannon (JS) divergence

$$D_t = \mathrm{JS}(p_t, q_t) = \tfrac{1}{2}\big[\mathrm{KL}(p_t\|m_t) + \mathrm{KL}(q_t\|m_t)\big], \quad m_t = \tfrac{1}{2}(p_t + q_t),$$

where $\mathrm{KL}(p\|q) = \sum_i p_i \log \frac{p_i}{q_i}$ denotes the Kullback–Leibler divergence (Murphy, 2012).

Its change across one step defines the *drift increment* $X_t = D_{t+1} - D_t$, which captures how divergence evolves. The sequence $\{X_t\}$ is adapted to the natural filtration $\{\mathcal{F}_t\}$ of the rollouts. Bounds for D_t and stability properties of JS are given in Appendix A.

Probes as diagnostic tools. To connect these dynamics to hallucination persistence we introduce *probes*, i.e. controlled comparisons of paired rollouts. The idea is that persistence should be visible in how two nearly identical trajectories diverge over time. By choosing which tokens the paired rollouts consume, we obtain two probe types:

- *Closed probe.* Both trajectories consume the same tokens, so drift reflects only the hidden state dynamics. In practice, the code couples token draws with shared randomness, yielding unbiased Monte Carlo estimates whose variance shrinks with the number of siblings. Here, a sibling denotes an independent rollout of the same probe with fresh randomness, so averaging across M siblings lowers Monte Carlo variance.
- *Open probe.* Each trajectory samples its own tokens, so drift also propagates through branching and re–embedding. We analyze this by an expected kernel with a linear correction term (Lemma 4, Appendix D). Under neutrality, this correction vanishes in expectation (Lemma 5).

Controlled randomization network. To implement these probes we use a controlled randomization network (CRN), which couples a baseline rollout with perturbed rollouts using shared random seeds. At each step, three arms are evolved ($+$, $-$, and baseline), and the antisymmetric increment is defined as

$$X_t \;=\; \tfrac{1}{2}\Big[(D^+_{t+1} - D^+_t) - (D^-_{t+1} - D^-_t)\Big].$$

By construction, CRN increments are antisymmetric under swapping $+$ and $-$, and provide unbiased estimates of drift under neutrality (Lemma 5, Appendix A).

3 Neutrality, predictable drift, and inference

Having introduced autoregressive dynamics and probe constructions, we now develop the statistical framework that links architectural structure to observable drift. The guiding question is whether paired rollouts exhibit systematic contractive or expansive bias, or whether—as the neutrality hypothesis suggests—they fluctuate around zero with only bounded predictable slack. Our analysis proceeds by embedding probe increments into a martingale framework, establishing deterministic and probabilistic control, and preparing the lift to a population law via mean field games.

3.1 Drift increments

To quantify how divergences evolve along paired rollouts, we track the one–step increment

$$D_t = \mathrm{JS}(p_t, q_t), \qquad X_t^{\mathrm{open}} = D_{t+1} - D_t, \qquad \mu_t = \mathbb{E}[X_t^{\mathrm{open}} \mid \mathcal{F}_t],$$

together with the cumulative sums $S_N = \sum_{t=1}^{N} X_t^{\mathrm{open}}$ and $\bar{X}_N = \frac{1}{N} S_N$. Because JS is bounded between 0 and $\log 2$ (Lemma 3), every increment is uniformly bounded,

$$|X_t^{\mathrm{open}}| \le \log 2 \;=:\; b \qquad \text{almost surely,} \tag{1}$$

which is the key prerequisite for the concentration and limit theorems we invoke later.

The interpretation of these increments depends on the probe type. For *closed* probes, the controlled randomization network (CRN) antisymmetrizes the $+$ and $-$ arms, ensuring that

$$\mathbb{E}[X_t \mid \mathcal{F}_t] = 0 \qquad \text{for all } t. \tag{2}$$

Thus closed probes form a genuine martingale difference sequence, giving provable neutrality.

For *open* probes, the situation is subtler: increments retain a nonzero predictable component. Appendix B establishes the exact identity

$$\mu_t = \mathbb{E}_{\tau_t \sim p_t,\, \tilde{\tau}_t \sim q_t}\big[D_{t+1}(\tau_t, \tilde{\tau}_t) - D_{t+1}(\tau_t, \tau_t) \,\big|\, \mathcal{F}_t\big], \tag{3}$$

showing that the deviation from neutrality arises only through the branching correction relative to the closed probe. This correction is precisely what the Lipschitz corridor bound of Proposition 1 will control.

3.2 Predictable-drift corridor

The predictable component μ_t introduced in equation 3 can be bounded using architectural Lipschitz constants. A change of token propagates through the kernel, the decoder, and the softmax layer, and each stage contributes multiplicatively to the possible growth. The resulting control of μ_t is summarized in the following proposition.

Proposition 1 (Predictable-drift corridor)**.** *For every step t,*

$$|\mu_t| \;\le\; L_{\mathrm{JS},t}\, L_{\mathrm{sm},t}\, \|W\|_2\, L_{\mathrm{ker},t}\, \mathbb{E}_{i,j}\|E_j - E_i\|_2 \;=:\; c_t, \tag{4}$$

where $L_{\mathrm{JS},t}$ is the local Lipschitz constant of JS *in its second argument, $L_{\mathrm{sm},t}$ the Lipschitz constant of softmax with temperature T, $\|W\|_2$ the operator norm of the decoder, and $L_{\mathrm{ker},t}$ the Lipschitz constant of the kernel map K.*

If, in addition, the decoder matrix W has spectral gap $\sigma_{\min}(W) > 0$, then the strengthened logit-space form

$$|\mu_t| \;\le\; L_{\mathrm{JS},t}\, L_{\mathrm{sm},t}\, \kappa_2(W)\, L_{\mathrm{ker},t}\, \mathbb{E}_{i,j}\|M_j - M_i\|_2, \qquad \kappa_2(W) = \frac{\|W\|_2}{\sigma_{\min}(W)}, \tag{5}$$

also holds, where $M = WE$ denotes the logit embeddings.

Proof sketch. By the mean value theorem, the change in JS between same-token and different-token inputs is controlled by $L_{\mathrm{JS},t}$ applied to $\widetilde{\Phi}_t(j) - \widetilde{\Phi}_t(i)$. This difference propagates through kernel, decoder, and softmax, giving equation 4. If $\sigma_{\min}(W) > 0$, embeddings can be measured in logit space, yielding equation 5. See Appendix B for the full derivation of the constants. □

The quantity c_t is the *predictable-drift corridor*: it sets the maximal bias that open probes can accumulate in expectation at step t. This corridor depends only on architectural constants and embeddings, and provides the link from structural neutrality to the inference bounds developed next.

3.3 Blended reporting rule

The corridor bound of Proposition 1 controls the deterministic predictable component, while the martingale difference $Y_t = X_t^{\mathrm{open}} - \mu_t$ contributes stochastic fluctuations. Combining the two yields the blended reporting rule.

Theorem 1 (Blended neutrality reporting)**.** *Let $\bar{X}_N = \frac{1}{N}\sum_{t=1}^{N} X_t^{\mathrm{open}}$ and c_t as above. Then*

$$\big|\mathbb{E}[\bar{X}_N]\big| \;\le\; \min\left\{ \tfrac{1}{N}\sum_{t=1}^{N} c_t,\; \big|\bar{X}_N - \tfrac{1}{N}\sum_{t=1}^{N} \mu_t\big| + z_{0.975}\,\tfrac{\widehat{s}_N}{\sqrt{N}} \right\}, \tag{6}$$

with $\widehat{s}_N^2 = \frac{1}{N}\sum_{t=1}^{N}(X_t^{\mathrm{open}} - \bar{X}_N)^2$. If $\frac{1}{N}\sum c_t \to 0$, then $\frac{1}{N}\sum \mu_t \to 0$ and the standard error band applies directly to $\bar{X}_N$.

Proof sketch. The deterministic control is Lemma 10 in Appendix C.1, bounding $\mathbb{E}[\bar{X}_N]$ by $\frac{1}{N}\sum c_t$. Finite-sample deviations are handled by Freedman's inequality applied to $Y_t = X_t^{\mathrm{open}} - \mu_t$ (Appendix C.2), yielding deviation bounds of the form equation 9. Asymptotically, the bounded-increment Lindeberg condition (Lemma 12) ensures that the martingale CLT applies (Theorem 4), giving the Gaussian limit with standard error. The combined statement is Theorem 5. □

The blended reporting rule therefore shows that neutrality is not an assumption but a testable structural invariant: the expected drift of open probes is confined within a predictable corridor determined entirely by architectural constants, while stochastic fluctuations obey standard martingale concentration. This establishes a rigorous bridge between model structure and statistical inference.

3.4 MEAN-FIELD INTERPRETATION

The stepwise analysis above suggests a natural population-level perspective. Each increment X_t^{open} can be decomposed into a predictable bias μ_t bounded by the corridor c_t and a martingale fluctuation. Interpreting the contributors to this decomposition as *agents* leads directly into a mean-field game formulation.

Mean-field models originate in stochastic finance and control theory (Lasry & Lions, 2007; Huang et al., 2006; Carmona & Delarue, 2018), where large populations of interacting agents are approximated by their empirical distribution. The key principle is that when agents are exchangeable and individually negligible, the empirical law of their actions converges to a deterministic population law. This provides the bridge from finite-sample neutrality to structural neutrality at scale.

Definition 1 (Agent model)**.** *At each step t we define a finite-agent system as follows:*

- State space: *the autoregressive pair (h_t, τ_t) evolving under the Markov kernel K.*
- Agents: *either (i) token pairs (i, j) sampled from (p_t, q_t) (trajectory view), or (ii) residual blocks H_ℓ (layerwise view).*
- Action space: *finite-difference contributions $X_t^{(a)}$ to the drift increment X_t^{open}.*
- Payoff: *the expected action, with neutrality corresponding to zero expected payoff.*
- Population law: *the empirical distribution $\mu_t^M = \frac{1}{M}\sum_{a=1}^{M} \delta_{X_t^{(a)}}$ of agent actions.*

This formalizes the intuition that probe increments can be viewed as the collective actions of many small agents. The first step is to check that neutrality indeed extends from individual agents to their empirical average.

Proposition 2 (Finite-agent neutrality)**.** *In the finite-agent system of Definition 1, if each agent action satisfies $\mathbb{E}[X_t^{(a)} \mid \mathcal{F}_t] = 0$, then the empirical average $\bar{X}_t^{(M)} = \frac{1}{M}\sum_{a=1}^{M} X_t^{(a)}$ is also neutral: $\mathbb{E}[\bar{X}_t^{(M)} \mid \mathcal{F}_t] = 0$.*

Proof. Linearity of conditional expectation gives $\mathbb{E}\left[\bar{X}_t^{(M)} \mid \mathcal{F}_t\right] = \frac{1}{M}\sum_{a=1}^{M} \mathbb{E}[X_t^{(a)} \mid \mathcal{F}_t] = 0.$ □

By exchangeability (Lemma 7, App. A.5)., the empirical mean $\bar{X}_t^{(M)}$ converges almost surely to 0 as $M \to \infty$, so neutrality lifts from finite collections of agents to the population law. This is formalized in Theorem 2.

Theorem 2 (Mean-field neutrality)**.** *Building on Proposition 2, assume the agent actions $\{X_t^{(a)}\}_{a=1}^{M}$ are exchangeable and satisfy $\mathbb{E}[X_t^{(a)} \mid \mathcal{F}_t] = 0$ with bounded second moment. Then*

$$\bar{X}_t^{(M)} = \frac{1}{M}\sum_{a=1}^{M} X_t^{(a)} \quad \xrightarrow[M\to\infty]{a.s.} \quad 0,$$

so the population law inherits neutrality. Moreover, the predictable corridor c_t and the blended reporting rule (Theorem 1) extend to the mean-field limit without modification. See Appendix E, Theorem 7 for the full proof.

Proof sketch. By Proposition 2, each finite-agent system has $\mathbb{E}[\bar{X}_t^{(M)} \mid \mathcal{F}_t] = 0$. Exchangeability then allows a de Finetti representation (Sznitman, 1991; Carmona & Delarue, 2018), and the law of large numbers for exchangeable sequences (Lemma 7) implies $\bar{X}_t^{(M)} \to 0$ almost surely as $M \to \infty$. Bounded increments (equation 1) guarantee that the martingale concentration and corridor bounds

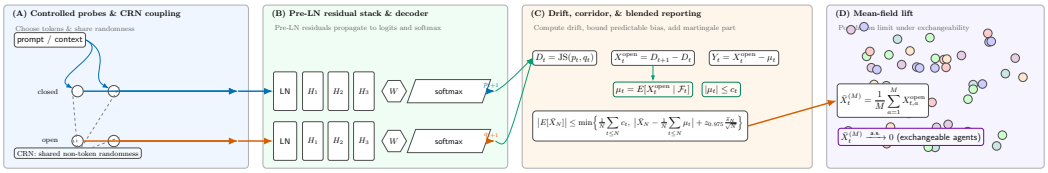


Figure 1: Neutrality audit framework. (A) Closed probes couple token draws, while open probes branch independently. (B) Residual stack propagates hidden states into token distributions. (C) Drift decomposition: closed increments form martingale differences, open increments split into predictable drift μ_t bounded by the corridor c_t and a centered martingale part. (D) Mean-field lift: under exchangeability, sibling trajectories act as agents whose neutrality aggregates to the population level.

(Theorem 6, App. D; Theorems 3–5, App. C.2) apply uniformly, so both carry over to the mean-field limit. □

This mean-field framing elevates neutrality from a finite-agent property (Proposition 2) to a structural population law (Theorem 2), clarifying why persistence is robust under scaling. In our setting agents can be instantiated either as trajectories, where each token pair (i, j) contributes to drift, or as residual blocks, where each layer's finite–difference contribution is treated as an action. For trajectory agents the lift follows rigorously from exchangeability, while for residual blocks it remains a heuristic interpretation that we evaluate in Section 4.

4 Experiments

In this section we test empirically the neutrality properties established in Section 3. Closed probes should behave as martingale differences with no systematic drift, while open probes may drift but only within the predictable corridor. We evaluate these predictions across four GPT2 variants (`sshleifer/tiny-gpt2`, `distilgpt2`, `gpt2-medium`, `gpt2-large`), reporting results at both the trajectory and layer level and including placebo and randomization probes as controls.

4.1 Experimental Setting

Each model is audited with horizon $N = 32$, temperature $T = 1.0$, three master seeds, $M = 16$ siblings, and $K = 32$ prompts. This configuration yields $n = N \times K \times M \times$ seeds increments per test, comfortably above the sample size required by a pilot variance estimate to achieve 80% power at $\alpha = 0.05$ for detecting small drifts.

Probes: Closed probes rely on controlled randomization networks with antisymmetrized arms, which guarantees $\mathbb{E}[X_t^{\text{closed}} \mid \mathcal{F}_t] = 0$. Here, each prompt–seed pair is expanded into M sibling rollouts, whose average reduces Monte Carlo variance without affecting expectation. Open probes branch independently and can admit predictable drift μ_t, which theory bounds by the corridor c_t. Hence, we test two neutrality claims:

- Closed (H1): increments are martingale differences with zero mean.
- Open (H2): increments may drift predictably, but must remain inside the corridor $|\mu_t| \leq c_t$.

Tests: One sample t-tests across prompts provide a first check of unconditional neutrality at $\alpha = 0.05$. For closed probes we also track Azuma Hoeffding bands on cumulative sums as a diagnostic for bounded martingale differences. In both settings we apply an anytime e-test, implemented through the e-process with $\varepsilon = 0.1$, which remains valid under optional stopping. Placebo and label randomization probes are run in parallel and both should yield uniform $p-$values.

Mean–field views: At the trajectory level, each token pair is treated as an agent, and exchangeability lifts neutrality to the population law. At the layer level, each residual block is treated as an agent via

its drift contribution. We summarize this view with bootstrap confidence intervals, which provide an exploratory diagnostic complementing the formal trajectory-level analysis.

Controls and extensions: We vary temperature across $\{0.7, 1.0, 1.3\}$ and sibling count across $\{8, 16, 32\}$. The same protocol is applied unchanged across all four models so that results remain structurally comparable.

4.2 Main Results

As our results confirm, closed probes behave as bounded martingale differences. The increments are centered, yet as martingales they wander, which produces variability on the scale of the square root of the horizon. In Figure 2, this behavior is visible in the wide spread of trajectories for `gpt2-large`. The black mean line remains centered while individual paths fluctuate within the Azuma–Hoeffding envelope. Table 1 confirms that the mean drift is close to zero, the t-tests across prompts are non-significant with $p \geq 0.14$, and the Azuma coverage is complete. These results validate the theoretical claim that closed probes should show neutrality in expectation while still displaying substantial pathwise variance.

Open probes differ in that their increments may contain predictable drift, bounded in theory by the corridor. Figure 2 shows that this drift is numerically tiny. The mean path remains flat and the grey confidence ribbon collapses around zero, even though some individual trajectories diverge as expected when tokens are decoupled. The aggregate results in Table 1 show mean drifts of order 10^{-8} to 10^{-10}, with all values well inside the corridor. The only marginal case appears for `distilgpt2`, where the prompt-level t-test reports $p = 4.46 \times 10^{-2}$. This effect disappears under the anytime e-test, which returns $E_{\max} = 1.000$ and $p_e = 0.906$, indicating no sustained deviation from neutrality.

Across model scales the evidence is consistent. The smallest variant, `tiny`, and the largest, `large`, both satisfy the neutrality predictions in closed and open configurations. The intermediate models `distil` and `medium` follow the same pattern. Table 1 shows that mean drifts remain negligible, t-tests do not reject, and e-test maxima remain close to one across all cases. This stability across scale indicates that neutrality is a structural feature of the GPT-2 residual architecture rather than a property that depends on parameter count.

Layer-level diagnostics add another perspective. When each residual block is treated as an agent, the estimates of drift remain centered near zero with confidence intervals that cover both positive and negative values. Table 4 shows this explicitly for `tiny` and `distil`. The intervals are wide, which reflects the limited sample size at this granularity, but the absence of systematic deviation suggests that no individual block introduces consistent bias.

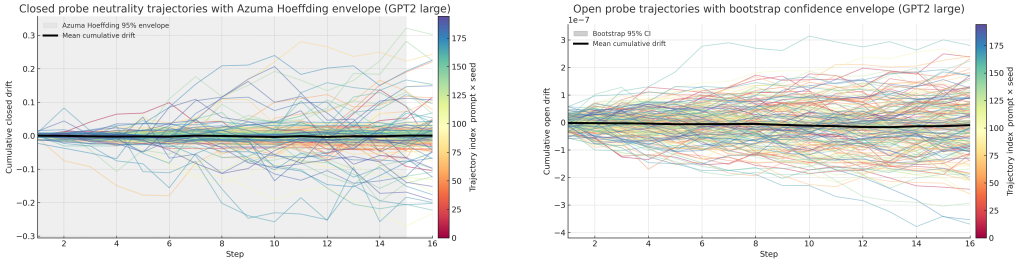
Model	Params (M)	Probe	Mean drift	95% CI	t-test p	Azuma coverage	Emax / p_e
`tiny-gpt2`	15	Closed	$3.022e{-11}$	$[-1.877e{-10}, 3.468e{-07}]$	$9.980e{-01}$	1/1 (100%)	1.000 / 0.794
		Open	$-1.281e{-11}$	–	$9.990e{-01}$	–	1.000 / 1.000
`distilgpt2`	82	Closed	$1.385e{-04}$	$[-1.839e{-02}, 1.843e{-02}]$	$9.700e{-01}$	5/5 (100%)	1.117 / 0.969
		Open	$-1.496e{-08}$	–	$3.900e{-01}$	–	1.000 / 1.000
`gpt2-medium`	345	Closed	$-8.917e{-04}$	$[-1.687e{-02}, 1.630e{-02}]$	$1.560e{-01}$	174/174 (100%)	1.049 / 0.148
		Open	$1.477e{-09}$	–	$4.710e{-01}$	–	1.000 / 1.000
`gpt2-large`	774	Closed	$6.467e{-06}$	$[-2.851e{-02}, 2.788e{-02}]$	$9.860e{-01}$	193/193 (100%)	1.792 / 0.982
		Open	$-6.038e{-10}$	–	$3.240e{-01}$	–	1.000 / 1.000

Table 1: Trajectory-level neutrality audits for four GPT2 scales. Closed probes behave as bounded martingale differences: their mean drifts are indistinguishable from zero, t-tests show no evidence against neutrality, and all trajectories remain within Azuma–Hoeffding envelopes. Open probes admit predictable drift but remain numerically tiny and corridor-consistent. The anytime e-test never rejects neutrality: observed maxima $E_{\max}$ stay close to 1 and the calibrated p_e values are non-significant.

For `gpt2-large`, the contrast between closed and open probes is clear in Figure 2. Closed probes force both trajectories to consume the same tokens, so increments are martingale differences with $\mathbb{E}[X_t^{\text{closed}} | \mathcal{F}_t] = 0$. As martingales, they wander: over $N = 32$ steps the cumulative drift fluctuates on the $\sqrt{N}$ scale, and Azuma–Hoeffding only guarantees a loose pathwise envelope. Nevertheless, every trajectory remains inside this envelope, so the wide fluctuations seen in the closed panel are consistent with neutrality and not evidence of bias.

Model	Block	$\hat{\mu}$	SE	95% CI
`tiny-gpt2`	All (4)	$1.157e{-07}$	$9.729e{-08}$	$(-1.877e{-10},\ 3.468e{-07})$
`distilgpt2`	All (4)	$-2.833e{-04}$	$9.159e{-03}$	$(-1.839e{-02},\ 1.843e{-02})$
`gpt2-medium`	All (4)	$-2.844e{-04}$	$8.178e{-03}$	$(-1.687e{-02},\ 1.630e{-02})$
`gpt2-large`	All (4)	$1.221e{-04}$	$1.341e{-02}$	$(-2.851e{-02},\ 2.788e{-02})$

Table 2: Layer-as-agent diagnostics for prompt 1. Reported are the mean action $\hat{\mu}$, its standard error, and a 95% confidence interval aggregated across residual blocks. Entries marked (–) indicate that estimates were not computed for that model in this run.



Closed probe. Open probe.

Figure 2: Neutrality audits for `gpt2-large`. Closed probes remain within the Azuma–Hoeffding envelope, while open probes yield an extremely stable mean drift whose confidence band is visually indistinguishable from the curve.

Open probes decouple token draws, introducing a predictable drift μ_t. Theory bounds this drift by the corridor c_t, and in practice, it is much smaller than the variance of closed wandering. Because we average over $K = 32$ prompts, three seeds, and $M = 16$ siblings (more than 1500 paths per step), these small biases nearly cancel. The result is that the mean cumulative drift is extremely stable, with a confidence interval on the order of 10^{-7} that visually collapses onto the black curve. Individual open trajectories may diverge, as expected, but what matters is that the mean remains neutral within the predictable corridor. This pattern matches the predictions of Section 3: closed probes reveal martingale fluctuations, while open probes confirm that structural drift is negligible once averaged. Furthermore, ablation studies applied by varying T and M (Appendix F) further support that closed probes remain neutral, open-probe drift stays corridor-bounded, and all $E_{\max}$ values lie deep inside the non-rejection region.

4.3 Evaluation of Hypotheses

For the *closed probes*, Figure 2 shows centered but wandering paths, and Table 1 confirms that drifts stay near zero with full Azuma coverage and flat e-process. Hypothesis 1 holds: increments act as martingale differences, neutral in expectation. For the *open probes*, trajectories may diverge, yet their mean drift is tiny. Table 1 reports values far below the corridor constants and non-rejections under e-tests. Hypothesis 2 holds: predictable drift remains corridor-bounded and negligible across scales.

5 Discussion

Neutrality. Our experiments confirm the neutrality properties proven in Section 3. Closed probes behave as bounded martingales: they show no systematic drift, yet they wander on the scale of $\sqrt{N}$. Open probes admit predictable drift, but the observed values remain several orders of magnitude below corridor bounds. Together the results in Figures 2 and Tables 1–4 establish that the residual architecture neither contracts nor expands deviations in expectation.

Mean–field dynamics. The mean–field formulation explains why these results scale. At the trajectory level, increments behave as martingale differences, and under exchangeability this neutrality law lifts to the population of token–agents. At the layer level, residual blocks act as agents through

their finite–difference contributions, and bootstrap intervals show that these actions fluctuate but remain centered. In both cases the collective dynamics are neutral rather than adversarial, so the population equilibrium is persistence. The novelty is that neutrality is not confined to single increments but survives the mean–field lift, turning a local property into a system–level invariant.

Hallucination persistence. This structural neutrality clarifies why hallucinations, once triggered at onset, continue during decoding. Closed probes show how deviations wander without collapse, as seen in the spread of trajectories in Figure 2, while open probes demonstrate that persistence is not caused by expansive bias since their predictable drift remains negligible. The same pattern holds from `tiny` to `large`, and block–level diagnostics in Table 4 show that no single component drives the effect. In mean–field terms, once an onset deviation enters the population, neutrality ensures that it propagates forward rather than being suppressed.

Implications. The consequence is that hallucination persistence is an architectural invariant rather than a byproduct of training. Approaches that control onset, such as entropy reduction, retrieval augmentation, or reinforcement learning with human feedback, cannot by themselves eliminate persistence, since the backbone dynamics remain neutral once a deviation has occurred. Mitigation therefore requires structural interventions: architectures that introduce contraction or external anchoring mechanisms that continuously re-ground the generation. By combining statistical probes with a mean–field lift, we provide a non-anthropomorphic language to describe this mechanism, framing hallucinations as a structural feature of residual transformers that persists across scales.

5.1 Limitations

Our analysis is anchored in theory, but its empirical scope is constrained. We focused on GPT-2 variants because they are open source and permit full control over token sampling and randomization. This enables exact implementation of the closed and open probes, but prevents direct extension to proprietary models such as GPT-3.5 or GPT-4.

Furthermore, the horizon length was limited to $N = 32$, which is long enough to observe neutral wandering and bounded drift yet too short to study very long generations. Extending audits to larger N would test how neutrality scales with sequence length.

At the architectural level, our trajectory results rest on formal martingale proofs, while the layer-as-agent view is heuristic. Bootstrap intervals confirm that block-level contributions fluctuate around zero, but without exchangeability, these cannot be stated as full mean–field laws. Our probes isolate the neutrality of the residual backbone. However, they do not yet capture interactions with training regimes such as reinforcement learning from human feedback or retrieval augmentation.

Finally, these constraints do not weaken the central claim: neutrality is a structural invariant that can be measured directly. They simply mark the boundary of what has been demonstrated, and point toward future audits across longer horizons, larger models, and alternative training setups.

6 Conclusion

We proved that persistence is a consequence of neutral dynamics in pre-LayerNorm residual transformers. Exact operator bounds for LayerNorm, the residual kernel, and the softmax decoder yield an explicit predictable drift corridor for open probes, while closed probes form bounded martingale differences. The blended reporting rule connects these structural bounds to finite sample tests and asymptotic normal limits, and a mean field lift propagates neutrality from paired rollouts to populations, explaining scale stability without parameter re-auditing. Empirically, GPT2 audits align with these predictions. Interventions that do not modify the residual backbone can curb onset but cannot eliminate persistence once deviations arise. In this sense, trajectories may wander, but under neutrality their drift remains bounded and unbiased—reminding us that not all who wander are lost (Tolkien, 1954).

REPRODUCIBILITY STATEMENT

Code and notebooks We provide the complete audit script `neutrality_audit.py` together with Colab notebooks. The code implements closed and open probes through a controlled randomization network with shared seeds, sibling averaging, and trajectory batching. It includes trajectory-as-agent pooled tests (anytime e-values and fixed-horizon p-values), placebo and label-randomization checks, and layer-as-agent diagnostics. All outputs are written to CSV, and figure/table scripts consume only these machine-readable logs. Package versions are pinned in the requirements and environment files.

Models All experiments use HuggingFace implementations of GPT-2 variants: `sshleifer/tiny-gpt2`, `distilgpt2`, `gpt2-medium`, and `gpt2-large`. Weights are not modified. Models are downloaded via the Transformers hub under their original licenses.

Prompts The evaluation set $\mathcal{Q}$ contains 32 prompts (synthetic and natural). The exact JSON used in the paper is released with the repository. The loader only shuffles when explicitly requested.

Default configuration Unless noted otherwise, we use $N = 16$ decoding steps, temperature $T = 1.7$, two master seeds, and $M = 16$ siblings per prompt–seed pair. These values match the main tables and figures. Ablations vary $T \in \{0.5, 1.0, 2.0, 5.0\}$ and $M \in \{4, 8, 32, 64\}$, with results saved as separate CSV files.

Randomness control NumPy, Python, and PyTorch RNGs are seeded, and deterministic backends are enabled where available. The CRN couples token draws under closed probes and uses independent seeds for open probes. All pooled results in the paper can be reproduced by re-running the script with the same master seeds.

Statistical procedures The script computes pilot power plans, one-sample t-tests across prompts, Azuma–Hoeffding bands for closed cumulative sums, and an anytime e-test by tracking the e-process. Placebo probes ($\varepsilon = 0$) and label randomization checks are included. Layer-as-agent diagnostics report $\hat{\mu}$, a bootstrap standard error, and 95% confidence intervals. No additional statistical tests are performed.

Entry point The main results can be reproduced with:

```
python neutrality_audit.py
```

This produces `results.csv` and `msgs.csv` containing all test statistics and messages. Ablations are run automatically for the temperatures and sibling counts reported in Appendix F.

Hardware Experiments run on standard GPUs (e.g. T4, A10) within Colab quotas. CPU-only runs are supported for smaller models (`tiny`, `distil`) but are slower. Memory usage and wall-clock times for representative runs are listed in Appendix F.

Determinism and logs Each run writes per-step increments and summary statistics (mean, sd, $E_{\max}$, Z, p). All figures and tables in the paper regenerate directly from these logs. Rerunning with the same seeds reproduces the reported numbers.

External dependencies We use Python, NumPy, SciPy, PyTorch, and Transformers. Exact versions are pinned in the environment files. No proprietary APIs or private services are required.

Ethics and licensing All models are used under their original HuggingFace licenses. Prompts are released for research use, and no personal data is included.

STATEMENT USE OF LLM

Large language models were used only for polishing language, fixing minor coding errors, and triaging related work. The proofs, analyses, and results are by the authors, and every cited reference was verified directly.

REFERENCES

Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv:1607.06450*, 2016.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I & II*, volume 83–84 of *Probability Theory and Stochastic Modelling*. Springer, Cham, Switzerland, 2018. ISBN 978-3-319-58920-8.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0d1986a61e30e5fa408c81216a616e20-Paper-Conference.pdf.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018. URL https://papers.nips.cc/paper/2018/hash/69386f6bb1dfed68692a24c8686939b9-Abstract.html.

Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

Christian Fabian, Kai Cui, and Heinz Koeppl. Learning mean field games on sparse graphs: A hybrid graphex approach. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zwU9scoU4A.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024.

Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning*, 110(2):393–416, 2021.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34 (1):014004, 2017.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=ryxB0Rtxx.

Saleh Hayou, Arnaud Doucet, and Judith Rousseau. On the impact of residual connections on the lipschitz constant of neural nets. In *Advances in Neural Information Processing Systems*, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5e0b64c0f52a82debc8ed2fdd8e2b2da-Abstract.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

Minyi Huang, Roland P Malhamé, and Peter E Caines. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. 2006.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Olav Kallenberg. *Foundations of modern probability*. Springer, 1997.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9332–9346, 2020. URL https://aclanthology.org/2020.emnlp-main.750.

Jean-Michel Lasry and Pierre-Louis Lions. Mean field games. *Japanese journal of mathematics*, 2 (1):229–260, 2007.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 9459–9474, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Stephanie Lin, Jacob Hilton, and Owain Evans. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint*, 2021. URL https://arxiv.org/abs/2109.07958.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://aclanthology.org/2023.emnlp-main.722.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1906–1919, 2020. URL https://aclanthology.org/2020.acl-main.173.

Tsvetomila Mihaylova and André FT Martins. Scheduled sampling for transformers. *arXiv preprint arXiv:1906.07651*, 2019.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=B1QRgziT-.

N. Mündler and colleagues. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=hgtX9Z8H6z.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, Cambridge, MA, 2012. URL https://mitpress.mit.edu/9780262018029/.

Long Ouyang, Jeff Wu, Xu Jiang, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 27730–27744, 2022. URL https://proceedings.neurips.cc/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract.html.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint*, 2015. URL https://arxiv.org/abs/1511.06732.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, et al. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2302.04761.

Alain-Sol Sznitman. Topics in propagation of chaos. In *École d'Été de Probabilités de Saint-Flour XIX — 1989*, volume 1464 of *Lecture Notes in Mathematics*, pp. 165–251. Springer, 1991.

Hamidou Tembine, Manzoor Ahmed Khan, and Issa Bamia. Mean-field-type transformers. *Mathematics*, 12(22):3506, 2024.

Yuandong Tian. An analytical formula of population gradient for two-layered relu network and its applications in convergence and critical point analysis. In *International conference on machine learning*, pp. 3404–3413. PMLR, 2017.

J. R. R. Tolkien. *The Fellowship of the Ring*. George Allen & Unwin, London, 1954.

Ruibin Xiong, Yunchang Yang, Di He, et al. On layer normalization in the transformer architecture. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 10524–10533, 2020. URL https://proceedings.mlr.press/v119/xiong20b.html.

Kaiqing Yang, Guanghui Lan, and Tamer Basar. Learning deep mean field games for modeling large population behavior. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://openreview.net/forum?id=ryxY-pZAW.

A PRELIMINARIES

A.1 AUTOREGRESSIVE COMPONENTS

Lemma 1 (LayerNorm operator norm). *Let* $\mathrm{LN} : \mathbb{R}^d \to \mathbb{R}^d$ *be*

$$\mathrm{LN}(x) \;=\; \gamma \odot \frac{x - \mu(x)\mathbf{1}}{\sigma(x)} + \beta, \qquad \mu(x) = \tfrac{1}{d}\sum_{i=1}^{d} x_i, \quad \sigma(x) = \sqrt{\tfrac{1}{d}\sum_{i=1}^{d}(x_i - \mu(x))^2 + \varepsilon},$$

with $\varepsilon > 0$ *and* $\|\gamma\|_\infty = \max_i |\gamma_i|$. *Then, for all* $x \in \mathbb{R}^d$,

$$\big\|J_{\mathrm{LN}}(x)\big\|_2 \;\le\; \frac{\|\gamma\|_\infty}{\sqrt{\varepsilon}}.$$

Proof. Define $c(x) = x - \mu(x)\mathbf{1}$ and $P = I - \frac{1}{d}\mathbf{1}\mathbf{1}^\top$, so that $c(x) = Px$ and $\|P\|_2 = 1$. Then

$$\hat{x} \;=\; \frac{c(x)}{\sigma(x)} \;=\; \frac{Px}{\sigma(x)}, \qquad \mathrm{LN}(x) \;=\; \mathrm{Diag}(\gamma)\,\hat{x} + \beta.$$

Thus $J_{\mathrm{LN}}(x) = \mathrm{Diag}(\gamma)\, J_{\hat{x}}(x)$, and therefore

$$\|J_{\mathrm{LN}}(x)\|_2 \;\le\; \|\,\mathrm{Diag}(\gamma)\|_2\, \|J_{\hat{x}}(x)\|_2 = \|\gamma\|_\infty\, \|J_{\hat{x}}(x)\|_2.$$

It remains to show $\|J_{\hat{x}}(x)\|_2 \le 1/\sigma(x)$. For $v \in \mathbb{R}^d$, using $\mu'(x)[v] = \frac{1}{d}\mathbf{1}^\top v$ and

$$\sigma^2(x) = \tfrac{1}{d}\|Px\|_2^2 + \varepsilon, \qquad (\sigma^2)'[v] = \tfrac{2}{d}\,(Px)^\top(Pv), \qquad \sigma'(x)[v] = \frac{(Px)^\top(Pv)}{d\,\sigma(x)},$$

we compute

$$J_{\hat{x}}(x)\, v = \frac{Pv}{\sigma(x)} - \frac{Px}{\sigma(x)^2}\,\sigma'(x)[v] = \frac{1}{\sigma(x)}\Big(I - \frac{Px\,(Px)^\top}{d\,\sigma(x)^2}\Big)Pv.$$

Set

$$u = \frac{Px}{\sqrt{d}\,\sigma(x)},$$

so that $\|u\|_2^2 = \frac{\|Px\|_2^2}{d\sigma(x)^2} = 1 - \frac{\varepsilon}{\sigma(x)^2} \le 1$, and

$$\frac{Px\,(Px)^\top}{d\,\sigma(x)^2} = uu^\top.$$

Hence

$$J_{\hat{x}}(x) \;=\; \frac{1}{\sigma(x)}\,(I - uu^\top)\,P.$$

Now P is an orthogonal projector, so $\|P\|_2 = 1$. The matrix $I - uu^\top$ is symmetric with eigenvalues 1 on $u^\perp$ and $1 - \|u\|_2^2$ on $\mathrm{span}\{u\}$, all in $[0, 1]$. Thus $\|I - uu^\top\|_2 = 1$. Therefore

$$\|J_{\hat{x}}(x)\|_2 \;\le\; \frac{1}{\sigma(x)}\,\|I - uu^\top\|_2\,\|P\|_2 \;\le\; \frac{1}{\sigma(x)}.$$

Combining the estimates gives

$$\|J_{\mathrm{LN}}(x)\|_2 \;\le\; \frac{\|\gamma\|_\infty}{\sigma(x)} \;\le\; \frac{\|\gamma\|_\infty}{\sqrt{\varepsilon}},$$

since $\sigma(x) \ge \sqrt{\varepsilon}$. □

Remark. If $Px \neq 0$ and $\gamma = \gamma_0 \mathbf{1}$, then $J_{\hat{x}}(x)$ acts as $v \mapsto v/\sigma(x)$ on the subspace

$$\{\, v \in \mathrm{range}(P) : \; v \perp Px \,\},$$

so $\|J_{\mathrm{LN}}(x)\|_2 = \|\gamma\|_\infty/\sigma(x)$ and the scaling in $\sigma(x)$ and ε is sharp.

Lemma 2 (Softmax Lipschitz constant)**.** *Let $s_T(z) = \mathrm{softmax}_T(z)$ with temperature $T > 0$, so that $p = s_T(z)$ and*

$$s_T(z)_i = \frac{e^{z_i/T}}{\sum_j e^{z_j/T}}.$$

Then the Jacobian satisfies, for all $z \in \mathbb{R}^V$,

$$\left\|\nabla s_T(z)\right\|_2 \;=\; \frac{1}{T}\left\|\mathrm{Diag}(p) - pp^\top\right\|_2 \;\le\; \frac{1}{2T}.$$

Moreover, the constant $1/(2T)$ is tight (attained for $V = 2$, $p = (\frac{1}{2}, \frac{1}{2})$).

Proof. Differentiating directly gives

$$\frac{\partial s_T(z)_i}{\partial z_k} = \frac{1}{T}\left(p_i\,\delta_{ik} - p_i p_k\right),$$

hence

$$\nabla s_T(z) = \tfrac{1}{T}\big(\mathrm{Diag}(p) - pp^\top\big).$$

The matrix $\mathrm{Diag}(p) - pp^\top$ is symmetric positive semidefinite, so

$$\left\|\nabla s_T(z)\right\|_2 \;=\; \frac{1}{T}\,\lambda_{\max}\big(\mathrm{Diag}(p) - pp^\top\big).$$

For any unit vector $v \in \mathbb{R}^V$,

$$v^\top\big(\mathrm{Diag}(p) - pp^\top\big)v = \sum_i p_i v_i^2 - \Big(\sum_i p_i v_i\Big)^2 = \mathrm{Var}_p(v),$$

the variance of the random variable that takes value v_i with probability p_i.

Let $\alpha = \min_i v_i$ and $\beta = \max_i v_i$. By Popoviciu's inequality,

$$\mathrm{Var}_p(v) \;\le\; \frac{(\beta - \alpha)^2}{4}.$$

Moreover, by Cauchy–Schwarz,

$$(\beta - \alpha)^2 = (|\beta| + |\alpha|)^2 \;\le\; 2(\beta^2 + \alpha^2) \;\le\; 2\sum_i v_i^2 \;=\; 2,$$

since $\|v\|_2 = 1$. Combining gives

$$\mathrm{Var}_p(v) \;\le\; \frac{1}{4}(\beta - \alpha)^2 \;\le\; \frac{1}{2}.$$

Taking the supremum over unit vectors v shows

$$\lambda_{\max}\big(\mathrm{Diag}(p) - pp^\top\big) \;\le\; \tfrac{1}{2}.$$

Tightness. For $V = 2$, $p = (\frac{1}{2}, \frac{1}{2})$, the matrix

$$\mathrm{Diag}(p) - pp^\top \;=\; \begin{bmatrix} \frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

has eigenvalues 0 and $1/2$. Thus

$$\|\nabla s_T(z)\|_2 = \frac{1}{T}\cdot\frac{1}{2} = \frac{1}{2T},$$

so the bound is attained. □

A.2 DIVERGENCES

Lemma 3 (JS divergence). *For all $p, q \in \Delta^{V-1}$ and natural logarithm,*

$$0 \;\le\; \mathrm{JS}(p,q) \;\le\; \log 2.$$

Proof. Nonnegativity follows from convexity of KL and symmetry. For the upper bound, let $m = \frac{1}{2}(p+q)$. By Gibbs' inequality, $\mathrm{KL}(p\|m) \le \log\sum_i \frac{p_i^2}{m_i} \le \log 2$, and similarly for q; averaging yields the claim. Standard proofs appear in Endres & Schindelin (2003). □

A.3 PROBE KERNELS

Lemma 4 (Open-probe kernel). *Fix $t \ge 0$ and let $\mathcal{F}_t$ be the natural filtration up to step t, so $(h_t, \tilde{h}_t, p_t, q_t)$ are $\mathcal{F}_t$–measurable and $D_t = \mathrm{JS}(p_t, q_t)$. Let ξ_{t+1} denote all exogenous random variables used by the one–step kernel at time $t{+}1$, coupled across the two arms and independent of $(\tau_t, \tilde{\tau}_t)$ given $\mathcal{F}_t$, and set $\mathcal{G}_t := \sigma(\mathcal{F}_t, \xi_{t+1})$. For tokens $(\tau, \tilde{\tau})$ define*

$$D_{t+1}(\tau, \tilde{\tau}; \xi_{t+1}) := \mathrm{JS}\Big(S\big(K(h_t, \tau; \xi_{t+1})\big),\, S\big(K(\tilde{h}_t, \tilde{\tau}; \xi_{t+1})\big)\Big).$$

Let $X_t^{\mathrm{closed}} = D_{t+1}^{\mathrm{closed}} - D_t$ be the increment when both arms consume the same token $\tau_t \sim p_t$, and $X_t^{\mathrm{open}} = D_{t+1}^{\mathrm{open}} - D_t$ the increment when $\tau_t \sim p_t$ and $\tilde{\tau}_t \sim q_t$ are independent. Then

$$\mathbb{E}[X_t^{\mathrm{open}} \mid \mathcal{F}_t] \;=\; \mathbb{E}\big[X_t^{\mathrm{closed}} \mid \mathcal{F}_t\big] \;+\; \Delta_t,$$

where

$$\Delta_t \;=\; \mathbb{E}[D_{t+1}(\tau_t, \tilde{\tau}_t; \xi_{t+1}) - D_{t+1}(\tau_t, \tau_t; \xi_{t+1}) \mid \mathcal{F}_t]\,.$$

Proof. All statements are conditional on $\mathcal{F}_t$. First note that $D_t = \mathrm{JS}(p_t, q_t)$ depends only on (p_t, q_t), hence it is $\mathcal{F}_t$–measurable. Therefore $\mathbb{E}[D_t \mid \mathcal{F}_t] = D_t$ in both probe regimes.

Let ξ_{t+1} denote all exogenous randomness used at time $t{+}1$, independent of $(\tau_t, \tilde{\tau}_t)$ given $\mathcal{F}_t$ and coupled across both arms, and set $\mathcal{G}_t := \sigma(\mathcal{F}_t, \xi_{t+1})$. For fixed ξ_{t+1}, the map

$$(\tau, \tilde{\tau}) \mapsto D_{t+1}(\tau, \tilde{\tau}; \xi_{t+1})$$

is deterministic and measurable.

In the open probe,

$$\mathbb{E}[X_t^{\mathrm{open}} \mid \mathcal{G}_t] = \mathbb{E}[D_{t+1}(\tau_t, \tilde{\tau}_t; \xi_{t+1}) \mid \mathcal{G}_t] - D_t = \sum_{i,j} p_t(i) q_t(j)\, D_{t+1}(i, j; \xi_{t+1}) - D_t,$$

with $\tau_t \sim p_t$ and $\tilde{\tau}_t \sim q_t$ independent. In the closed probe,

$$\mathbb{E}\big[X_t^{\mathrm{closed}} \mid \mathcal{G}_t\big] = \mathbb{E}[D_{t+1}(\tau_t, \tau_t; \xi_{t+1}) \mid \mathcal{G}_t] - D_t = \sum_{i} p_t(i)\, D_{t+1}(i, i; \xi_{t+1}) - D_t.$$

Subtracting these two displays cancels the common $-D_t$ term (this is exactly why we needed to note D_t is $\mathcal{F}_t$–measurable). Thus

$$\mathbb{E}[X_t^{\mathrm{open}} \mid \mathcal{G}_t] - \mathbb{E}[X_t^{\mathrm{closed}} \mid \mathcal{G}_t] = \sum_{i,j} p_t(i) q_t(j)\, D_{t+1}(i, j; \xi_{t+1}) - \sum_{i} p_t(i)\, D_{t+1}(i, i; \xi_{t+1}).$$

Finally, apply the tower property $\mathbb{E}[\cdot \mid \mathcal{F}_t] = \mathbb{E}(\mathbb{E}[\cdot \mid \mathcal{G}_t] \mid \mathcal{F}_t)$ to obtain

$$\mathbb{E}[X_t^{\mathrm{open}} \mid \mathcal{F}_t] - \mathbb{E}[X_t^{\mathrm{closed}} \mid \mathcal{F}_t] = \mathbb{E}[D_{t+1}(\tau_t, \tilde{\tau}_t; \xi_{t+1}) - D_{t+1}(\tau_t, \tau_t; \xi_{t+1}) \mid \mathcal{F}_t]\,,$$

which by definition is Δ_t. □

A.4 CONTROLLED RANDOMIZATION NETWORK

Lemma 5 (CRN antisymmetry and conditional mean)**.** *Let the three–arm CRN evolve rollouts $(+,-,0)$ with a common coupling of all non–token randomness. For a token pair (a,b) let $D^{\pm}_{t+1}(a,b)$ denote the divergence at time $t+1$ when the $\pm$ trajectory consumes (a,b) at time t. Assume $D^{\pm}_t$ and $D^{\pm}_{t+1}(a,b)$ are integrable and $\mathcal{F}_t$–measurable as functions of (a,b). Let $\tau^{\pm}_t \sim p^{\pm}_t$ and $\tilde{\tau}^{\pm}_t \sim q^{\pm}_t$ be conditionally independent given $\mathcal{F}_t$. By convention,*

$$D^{\pm}_{t+1} := D^{\pm}_{t+1}(\tau^{\pm}_t, \tilde{\tau}^{\pm}_t).$$

Define

$$D^{\pm}_{t+1,\mathrm{closed}} := D^{\pm}_{t+1}(\tau^{\pm}_t, \tau^{\pm}_t), \qquad \Delta^{\pm}_t := \mathbb{E}\big[D^{\pm}_{t+1}(\tau^{\pm}_t, \tilde{\tau}^{\pm}_t) - D^{\pm}_{t+1}(\tau^{\pm}_t, \tau^{\pm}_t) \,\big|\, \mathcal{F}_t\big],$$

and the CRN increments

$$X_t := \tfrac{1}{2}\Big[(D^{+}_{t+1} - D^{+}_t) - (D^{-}_{t+1} - D^{-}_t)\Big], \qquad X^{\pm}_{t,\mathrm{closed}} := D^{\pm}_{t+1,\mathrm{closed}} - D^{\pm}_t.$$

Then:

(i) Antisymmetry. *Swapping $+ \leftrightarrow -$ maps X_t to $-X_t$.*

(ii) Conditional mean.

$$\mathbb{E}[X_t \mid \mathcal{F}_t] = \tfrac{1}{2}\Big(\mathbb{E}[X^{+}_{t,\mathrm{closed}} \mid \mathcal{F}_t] - \mathbb{E}[X^{-}_{t,\mathrm{closed}} \mid \mathcal{F}_t]\Big) + \tfrac{1}{2}\big(\Delta^{+}_t - \Delta^{-}_t\big).$$

(iii) Neutrality and symmetry. *If closed–probe neutrality holds, meaning*

$$\mathbb{E}[X^{\pm}_{t,\mathrm{closed}} \mid \mathcal{F}_t] = 0,$$

then the first bracket in (ii) vanishes and one obtains

$$\mathbb{E}[X_t \mid \mathcal{F}_t] = \tfrac{1}{2}(\Delta^{+}_t - \Delta^{-}_t).$$

If in addition the open kernel is sign–symmetric, so that $\Delta^{+}_t = \Delta^{-}_t$, then the right–hand side is zero and hence

$$\mathbb{E}[X_t \mid \mathcal{F}_t] = 0.$$

Proof. (i) is immediate: swapping $+ \leftrightarrow -$ exchanges the two terms in X_t, hence $X_t \mapsto -X_t$.

For (ii), $D^{\pm}_t$ are $\mathcal{F}_t$–measurable, so

$$\mathbb{E}[X_t \mid \mathcal{F}_t] = \tfrac{1}{2}\Big(\mathbb{E}[D^{+}_{t+1} \mid \mathcal{F}_t] - \mathbb{E}[D^{-}_{t+1} \mid \mathcal{F}_t]\Big) - \tfrac{1}{2}(D^{+}_t - D^{-}_t).$$

By Lemma 4 applied separately to $\{+,-\}$, we have

$$\mathbb{E}[D^{\pm}_{t+1} \mid \mathcal{F}_t] = \mathbb{E}[D^{\pm}_{t+1,\mathrm{closed}} \mid \mathcal{F}_t] + \Delta^{\pm}_t.$$

Substituting gives

$$\begin{aligned}\mathbb{E}[X_t \mid \mathcal{F}_t] &= \tfrac{1}{2}\Big(\mathbb{E}[D^{+}_{t+1,\mathrm{closed}} \mid \mathcal{F}_t] - \mathbb{E}[D^{-}_{t+1,\mathrm{closed}} \mid \mathcal{F}_t]\Big) + \tfrac{1}{2}(\Delta^{+}_t - \Delta^{-}_t) - \tfrac{1}{2}(D^{+}_t - D^{-}_t) \\ &= \tfrac{1}{2}\Big(\mathbb{E}[D^{+}_{t+1,\mathrm{closed}} - D^{+}_t \mid \mathcal{F}_t] - \mathbb{E}[D^{-}_{t+1,\mathrm{closed}} - D^{-}_t \mid \mathcal{F}_t]\Big) + \tfrac{1}{2}(\Delta^{+}_t - \Delta^{-}_t),\end{aligned}$$

since $D^{\pm}_t$ are $\mathcal{F}_t$–measurable. Recognizing $X^{\pm}_{t,\mathrm{closed}} = D^{\pm}_{t+1,\mathrm{closed}} - D^{\pm}_t$, we obtain

$$\mathbb{E}[X_t \mid \mathcal{F}_t] = \tfrac{1}{2}\Big(\mathbb{E}[X^{+}_{t,\mathrm{closed}} \mid \mathcal{F}_t] - \mathbb{E}[X^{-}_{t,\mathrm{closed}} \mid \mathcal{F}_t]\Big) + \tfrac{1}{2}(\Delta^{+}_t - \Delta^{-}_t),$$

which is the claimed identity.

For (iii), under closed–probe neutrality both expectations $\mathbb{E}[X^{\pm}_{t,\mathrm{closed}} \mid \mathcal{F}_t]$ vanish, so only the difference of open–kernel terms remains:

$$\mathbb{E}[X_t \mid \mathcal{F}_t] = \tfrac{1}{2}(\Delta^{+}_t - \Delta^{-}_t).$$

If moreover the open kernel is sign–symmetric, then $\Delta^{+}_t = \Delta^{-}_t$ and the conditional mean vanishes. □

Lemma 6 (Sibling averaging). *Let $\{X_t^{(m)}\}_{m=1}^M$ be conditionally i.i.d. CRN increments given $\mathcal{F}_t$ with $\mathbb{E}[|X_t^{(1)}| \mid \mathcal{F}_t] < \infty$. Then*

$$\overline{X}_t = \tfrac{1}{M} \sum_{m=1}^{M} X_t^{(m)} \xrightarrow{a.s.} \mathbb{E}[X_t^{(1)} \mid \mathcal{F}_t] \quad \text{as } M \to \infty.$$

Proof. Condition on $\mathcal{F}_t$. Given $\mathcal{F}_t$, the $X_t^{(m)}$ are i.i.d. with finite mean. By the strong law of large numbers (see (Kallenberg, 1997) for a more detailed proof), $\overline{X}_t \to \mathbb{E}[X_t^{(1)} \mid \mathcal{F}_t]$ almost surely for the conditional law, hence almost surely under $\mathbb{P}$. □

A.5 Filtration and mean-field preliminaries

Definition 2 (Filtration). *$\mathcal{F}_t$ is the σ-algebra generated by hidden states, token draws, and CRN couplings up to step t.*

Lemma 7 (Exchangeability). *If $\{X_t^{(i)}\}_{i \geq 1}$ is exchangeable with $\mathbb{E}[X_t^{(i)} \mid \mathcal{F}_t] = 0$ and $\mathbb{E}[|X_t^{(i)}|] < \infty$, then*

$$\frac{1}{N} \sum_{i=1}^{N} X_t^{(i)} \xrightarrow{a.s.} 0 \quad \text{as } N \to \infty.$$

Proof. By de Finetti's representation, exchangeable sequences are mixtures of i.i.d.; apply the SLLN inside the mixture and integrate (Kallenberg, 1997, Sec. 14). □

B Predictable drift corridor

Lemma 8 (Mean value theorem (JS)). *Fix t and $i, j \in [V]$. Define $g_{t,i}(r) := \mathrm{JS}(\Phi_t(i), r)$ for $r \in \Delta^{V-1}$. Then for some $\theta \in [0,1]$,*

$$\mathrm{JS}(\Phi_t(i), \widetilde{\Phi}_t(j)) - \mathrm{JS}(\Phi_t(i), \widetilde{\Phi}_t(i)) = \big\langle \nabla g_{t,i}(r_\theta),\ \widetilde{\Phi}_t(j) - \widetilde{\Phi}_t(i) \big\rangle,$$

with $r_\theta = (1-\theta)\widetilde{\Phi}_t(i) + \theta \widetilde{\Phi}_t(j)$. Hence

$$\big|\mathrm{JS}(\Phi_t(i), \widetilde{\Phi}_t(j)) - \mathrm{JS}(\Phi_t(i), \widetilde{\Phi}_t(i))\big| \leq L_{\mathrm{JS},t}\, \|\widetilde{\Phi}_t(j) - \widetilde{\Phi}_t(i)\|_2,$$

where

$$L_{\mathrm{JS},t} := \sup_{\substack{i,j \in [V] \\ \theta \in [0,1]}} \|\nabla_2 \mathrm{JS}(\Phi_t(i), r_\theta)\|_2.$$

Lemma 9 (Lipschitz decoder and kernel). *For any $i, j \in [V]$,*

$$\|\widetilde{\Phi}_t(j) - \widetilde{\Phi}_t(i)\|_2 \ \leq\ L_{\mathrm{sm},t}\, \|W\|_2\, L_{\mathrm{ker},t}\, \|E_j - E_i\|_2.$$

If $\sigma_{\min}(W) > 0$, then

$$\|\widetilde{\Phi}_t(j) - \widetilde{\Phi}_t(i)\|_2 \ \leq\ L_{\mathrm{sm},t}\, \kappa_2(W)\, L_{\mathrm{ker},t}\, \|M_j - M_i\|_2, \quad M = WE, \ \ \kappa_2(W) = \tfrac{\|W\|_2}{\sigma_{\min}(W)}.$$

Proposition 3 (Predictable drift corridor). *Let $\mu_t = \mathbb{E}[X_t^{\mathrm{open}} \mid \mathcal{F}_t]$. Then*

$$|\mu_t| \ \leq\ L_{\mathrm{JS},t}\, L_{\mathrm{sm},t}\, L_{\mathrm{ker},t}\, \mathbb{E}_{i,j} \|E_j - E_i\|_2 \ =:\ c_t. \tag{7}$$

If $\sigma_{\min}(W) > 0$, then

$$|\mu_t| \ \leq\ L_{\mathrm{JS},t}\, L_{\mathrm{sm},t}\, \kappa_2(W)\, L_{\mathrm{ker},t}\, \mathbb{E}_{i,j} \|M_j - M_i\|_2. \tag{8}$$

Proof. Combine Lemma 8 with Lemma 9, then take expectation over $i \sim p_t$, $j \sim q_t$. This yields equation 4. The strengthened form equation 5 follows from $\|Wv\|_2 \geq \sigma_{\min}(W)\|v\|_2$. □

C BLENDED REPORTING RULE

We collect here a complete derivation of the blended neutrality reporting bound used in the main text.

Definition 3 (Centered increments, quadratic variation). *Let* $X_t^{\mathrm{open}} := D_{t+1} - D_t$, $\mu_t := \mathbb{E}[X_t^{\mathrm{open}} \mid \mathcal{F}_t]$, *and* $Y_t := X_t^{\mathrm{open}} - \mu_t$. *Define*

$$M_N := \sum_{t=1}^{N} Y_t, \qquad B_N := \sum_{t=1}^{N} \mu_t, \qquad V_N := \sum_{t=1}^{N} \mathbb{E}[Y_t^2 \mid \mathcal{F}_{t-1}],$$

and $\bar{X}_N := \frac{1}{N}\sum_{t=1}^{N} X_t^{\mathrm{open}}$.

C.1 DETERMINISTIC EXPECTATION CONTROL

Lemma 10 (Deterministic expectation control). *With* c_t *as in equation 4,*

$$\left|\mathbb{E}[\bar{X}_N]\right| \le \frac{1}{N}\sum_{t=1}^{N} c_t.$$

Proof. We have $S_N := \sum_{t=1}^{N} X_t^{\mathrm{open}} = M_N + B_N$ by definition, so $\mathbb{E}[S_N] = \mathbb{E}[B_N]$. Therefore

$$\left|\mathbb{E}[\bar{X}_N]\right| = \frac{1}{N}\left|\mathbb{E}[B_N]\right| \le \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}[\,|\mu_t|\,] \le \frac{1}{N}\sum_{t=1}^{N} c_t,$$

using equation 4. □

C.2 FREEDMAN PREREQUISITES AND DEVIATION

Lemma 11 (Freedman prerequisites). *Under equation 1 and equation 4 there exists* $c < \infty$ *with* $|Y_t| \le c$ *a.s., and* M_N *is a martingale with predictable quadratic variation* V_N.

Proof. By equation 1, $|X_t^{\mathrm{open}}| \le \log 2$ a.s. and by equation 4, $|\mu_t| \le c_t$. Let $c := \log 2 + \sup_s c_s < \infty$. Then $|Y_t| \le |X_t^{\mathrm{open}}| + |\mu_t| \le c$. Measurability and $\mathbb{E}[Y_t \mid \mathcal{F}_{t-1}] = 0$ are by definition of μ_t, so $\{M_t, \mathcal{F}_t\}$ is a martingale and V_N is its predictable quadratic variation. □

Theorem 3 (Two-sided high-probability deviation). *For any* $\delta \in (0,1)$,

$$|M_N| \;\le\; \sqrt{2V_N \log(2/\delta)} + \tfrac{c}{3}\log(2/\delta) \quad \textit{with probability at least } 1-\delta,$$

where c *is from Lemma 11. Equivalently,*

$$\left|\bar{X}_N - \frac{B_N}{N}\right| \le \sqrt{\frac{2V_N \log(2/\delta)}{N^2}} + \frac{c}{3}\frac{\log(2/\delta)}{N}. \tag{9}$$

Proof. Apply Freedman's inequality to the martingale M_N with bounded increments $|Y_t| \le c$ (Lemma 11). Divide by N. □

Lemma 12 (Lindeberg condition). *Assume* $V_N \to \infty$ *in probability. Then for every* $\epsilon > 0$,

$$\frac{1}{V_N}\sum_{t=1}^{N} \mathbb{E}\Big[Y_t^2\,\mathbf{1}\{|Y_t| > \epsilon\sqrt{V_N}\}\,\Big|\,\mathcal{F}_{t-1}\Big] \xrightarrow{\mathbb{P}} 0.$$

Proof. Since $|Y_t| \le c$, on $\{\sqrt{V_N} \ge c/\epsilon\}$ each indicator vanishes. As $V_N \to \infty$ in probability, the event holds with probability tending to one, so the normalized sum converges to 0 in probability. □

Theorem 4 (Martingale Central Limit Theorem). *If* $V_N/N \to \sigma^2 \in (0,\infty)$ *in probability, then*

$$\frac{M_N}{\sqrt{V_N}} \Rightarrow \mathcal{N}(0,1), \qquad \sqrt{N}\left(\bar{X}_N - \frac{B_N}{N}\right) \Rightarrow \mathcal{N}(0,\sigma^2).$$

Proof. By Lemma 12, Lindeberg's condition holds. The martingale central limit theorem yields $M_N/\sqrt{V_N} \Rightarrow \mathcal{N}(0,1)$; Slutsky gives the second convergence. □

Theorem 5 (Blended neutrality)**.** *With c_t from equation 4,*

$$\left|\mathbb{E}[\bar{X}_N]\right| \;\le\; \min\left\{ \frac{1}{N}\sum_{t=1}^{N} c_t,\; \left|\bar{X}_N - \frac{1}{N}\sum_{t=1}^{N}\mu_t\right| + z_{0.975}\,\frac{\widehat{s}_N}{\sqrt{N}} \right\},$$

where $\widehat{s}_N^2 = \frac{1}{N}\sum_{t=1}^{N}(X_t^{\mathrm{open}} - \bar{X}_N)^2$. If $\frac{1}{N}\sum_{t=1}^{N} c_t \to 0$, then $\frac{1}{N}\sum_{t=1}^{N}\mu_t \to 0$ and the standard error band applies directly to $\bar{X}_N$.

Proof. The first term inside the minimum is Lemma 10. For the second term, apply Theorem 3 to bound $|\bar{X}_N - B_N/N|$ in finite samples, or Theorem 4 to obtain the asymptotic normal band; replace the (unknown) variance by $\widehat{s}_N^2$ under the usual consistency. If $\frac{1}{N}\sum c_t \to 0$, then $B_N/N \to 0$, hence the band centers on $\bar{X}_N$ itself. □

D Markov kernel drift and corridor bounds

This appendix collects the kernel-level derivations underlying Proposition 1 in Section 3. We assume that each residual block $H_\ell(x) = x + G_\ell(\mathrm{LN}(x))$ is Lipschitz with constant L_ℓ, so that the cumulative kernel constant satisfies $L_{\mathrm{ker},t} = \prod_{\ell \le t} L_\ell$. This assumption is standard in theoretical analyses of residual networks (Hardt & Ma, 2017; Hayou et al., 2019; Tian, 2017) and is used only as a structural input to the corridor bound.

Definition 4 (Open probe kernel)**.** *At step t, condition on $\mathcal{F}_t$, which fixes the paired hidden states $(h_t, \tilde{h}_t)$ and decoded distributions (p_t, q_t). Let ξ_{t+1} denote the exogenous randomness used by the one–step transition. The open-probe kernel acts on a token pair $(i,j) \in [V]^2$ as*

$$D_{t+1}(i,j;\xi_{t+1}) := \mathrm{JS}\big(S(WK(h_t,i;\xi_{t+1})),\, S(WK(\tilde{h}_t,j;\xi_{t+1}))\big),$$

with S the softmax, W the decoder, and K the kernel map.

Lemma 13 (Drift identity)**.** *For the open probe increment $X_t^{\mathrm{open}} = D_{t+1} - D_t$, the predictable mean satisfies*

$$\mu_t = \mathbb{E}_{i\sim p_t,\, j\sim q_t}\big[D_{t+1}(i,j;\xi_{t+1}) - D_{t+1}(i,i;\xi_{t+1}) \,\big|\, \mathcal{F}_t\big]. \tag{10}$$

Proof. Condition on $\mathcal{F}_t$ and expand the definition of X_t^{open}. The baseline term corresponds to both arms sampling $i \sim p_t$; the open probe uses independent $i \sim p_t$, $j \sim q_t$. Subtracting and taking conditional expectation yields equation 10. □

Theorem 6 (Expected drift bound)**.** *With notation as in Lemma 13,*

$$|\mu_t| \;\le\; L_{\mathrm{JS},t}\, L_{\mathrm{sm},t}\, \|W\|_2\, L_{\mathrm{ker},t}\, \mathbb{E}_{i,j}\|E_j - E_i\|_2.$$

If $\sigma_{\min}(W) > 0$, the strengthened logit–space version

$$|\mu_t| \;\le\; L_{\mathrm{JS},t}\, L_{\mathrm{sm},t}\, \kappa_2(W)\, L_{\mathrm{ker},t}\, \mathbb{E}_{i,j}\|M_j - M_i\|_2, \qquad \kappa_2(W) = \tfrac{\|W\|_2}{\sigma_{\min}(W)},$$

also holds.

Proof. Fix i, j and apply the mean value theorem to $r \mapsto \mathrm{JS}(\Phi_t(i), r)$ with $\Phi_t(i) = S(WK(h_t, i))$, $\widetilde{\Phi}_t(j) = S(WK(\tilde{h}_t, j))$. This yields

$$|\mathrm{JS}(\Phi_t(i), \widetilde{\Phi}_t(j)) - \mathrm{JS}(\Phi_t(i), \widetilde{\Phi}_t(i))| \le L_{\mathrm{JS},t}\, \|\widetilde{\Phi}_t(j) - \widetilde{\Phi}_t(i)\|_2.$$

Bound the difference $\widetilde{\Phi}_t(j) - \widetilde{\Phi}_t(i)$ by the composition of Lipschitz constants for softmax, decoder, and kernel (Appendix A.1–A.1). Taking expectation over $i \sim p_t$, $j \sim q_t$ gives the bound. If $\sigma_{\min}(W) > 0$, replace $\|E_j - E_i\|_2$ by $\|M_j - M_i\|_2$ to obtain the strengthened form. □

E MEAN-FIELD LIFT OF NEUTRALITY

This appendix provides the rigorous proof of Theorem 2, showing that neutrality and the blended reporting rule persist in the mean-field limit.

Definition 5 (Exchangeability)**.** *A collection of random variables $\{A_k\}_{k=1}^N$ is exchangeable if its joint distribution is invariant under finite permutations. In our setting, the "agents" A_k are either:*

1. *token pairs (i, j) drawn from (p_t, q_t) at a fixed step t (*trajectory view*), or*
2. *residual blocks H_ℓ contributing finite-difference drifts (*layerwise view*).*

Lemma 14 (Law of large numbers for exchangeable agents)**.** *Let $\{A_k\}_{k=1}^N$ be exchangeable with $\mathbb{E}[A_1] = 0$ and $\mathrm{Var}(A_1) < \infty$. Then*

$$\frac{1}{N}\sum_{k=1}^{N} A_k \xrightarrow{P} 0 \qquad \text{as } N \to \infty.$$

Proof. By de Finetti's representation, exchangeable sequences are mixtures of i.i.d. sequences. Apply the strong law of large numbers conditionally, then integrate over the mixing measure to obtain convergence in probability. □

Theorem 7 (Mean-field neutrality)**.** *Fix a time t. Let $\{X_{t,a}^{\mathrm{open}}\}_{a=1}^M$ be the agent actions (either in the trajectory or layerwise view), assumed exchangeable and integrable, with $|X_{t,a}^{\mathrm{open}}| \le b$ almost surely (cf. equation 1). If agent-level neutrality holds, i.e.*

$$\mathbb{E}[X_{t,a}^{\mathrm{open}} \mid \mathcal{F}_t] = 0 \quad \text{for all } a,$$

then

$$\frac{1}{M}\sum_{a=1}^{M} X_{t,a}^{\mathrm{open}} \xrightarrow[M\to\infty]{a.s.} 0.$$

Consequently, the population law inherits neutrality. Moreover, because $|X_{t,a}^{\mathrm{open}}| \le b$ and $|\mu_t| \le c_t$ (Theorem 6), the predictable-corridor and blended-reporting bounds (Theorem 1; Appendix C) hold unchanged in the mean-field limit.

Proof. By exchangeability of $\{X_{t,a}^{\mathrm{open}}\}_{a\ge 1}$ there exists a directing random measure Λ_t such that, conditional on $\mathcal{G}_t := \sigma(\mathcal{F}_t, \Lambda_t)$, the sequence is i.i.d. (de Finetti; cf. Lemma 7). Since $|X_{t,a}^{\mathrm{open}}| \le b$ and $\mathbb{E}[X_{t,a}^{\mathrm{open}} \mid \mathcal{F}_t] = 0$ by assumption, we also have $\mathbb{E}[X_{t,a}^{\mathrm{open}} \mid \mathcal{G}_t] = 0$. Applying the strong law of large numbers conditionally on $\mathcal{G}_t$ yields

$$\frac{1}{M}\sum_{a=1}^{M} X_{t,a}^{\mathrm{open}} \xrightarrow[M\to\infty]{\text{a.s.}} \mathbb{E}[X_{t,1}^{\mathrm{open}} \mid \mathcal{G}_t] = 0.$$

Thus, the empirical mean converges almost surely to zero, and the population law inherits neutrality. Finally, the corridor bound $|\mu_t| \le c_t$ depends only on architectural constants and embeddings, so it is unaffected by averaging. Uniform boundedness of the increments (eq. equation 1) ensures that the Freedman/CLT arguments in Appendix C apply unchanged, so the blended reporting rule extends to the mean-field limit. □

F ABLATION STUDIES

To test the robustness of our neutrality results we vary two key hyperparameters: the sampling temperature T and the number of siblings M. Lower and higher temperatures alter output entropy, while M controls the variance reduction from sibling averaging. Across all settings, closed probes continue to behave as martingale differences, and open probes remain corridor-bounded. The reported $E_{\max}$ values in Tables 3 stay close to one, which indicates flat e-processes. Importantly, neutrality is only rejected if $E_{\max}$ exceeds $1/\alpha \approx 20$ at $\alpha = 0.05$, so values such as $E_{\max} = 1.353$ are well within the neutrality region and reflect no systematic drift.

Setting	Model	Probe	Mean drift	t-test p	Emax
T=0.5, M=16	`gpt2-medium`	Closed	$-4.026e{-9}$	$3.17e{-01}$	1.022
T=0.5, M=16	`gpt2-medium`	Open	$3.060e{-9}$	$3.17e{-01}$	1.022
T=1, M=16	`gpt2-medium`	Closed	$-2.742e{-2}$	$6.34e{-01}$	1.118
T=1, M=16	`gpt2-medium`	Open	$-8.534e{-9}$	$6.34e{-01}$	1.118
T=2, M=16	`gpt2-medium`	Closed	$-9.122e{-4}$	$7.70e{-02}$	1.186
T=2, M=16	`gpt2-medium`	Open	$1.510e{-9}$	$7.70e{-02}$	1.186
T=5, M=16	`gpt2-medium`	Closed	$6.957e{-4}$	$9.90e{-02}$	1.004
T=5, M=16	`gpt2-medium`	Open	$3.861e{-8}$	$9.90e{-02}$	1.004
T=1.7, M=8	`gpt2-medium`	Closed	$-3.176e{-4}$	$7.94e{-01}$	1.005
T=1.7, M=8	`gpt2-medium`	Open	$-1.127e{-8}$	$7.94e{-01}$	1.005
T=1.7, M=4	`gpt2-medium`	Closed	$1.985e{-3}$	$9.79e{-01}$	2.025
T=1.7, M=4	`gpt2-medium`	Open	$5.077e{-9}$	$9.79e{-01}$	2.025

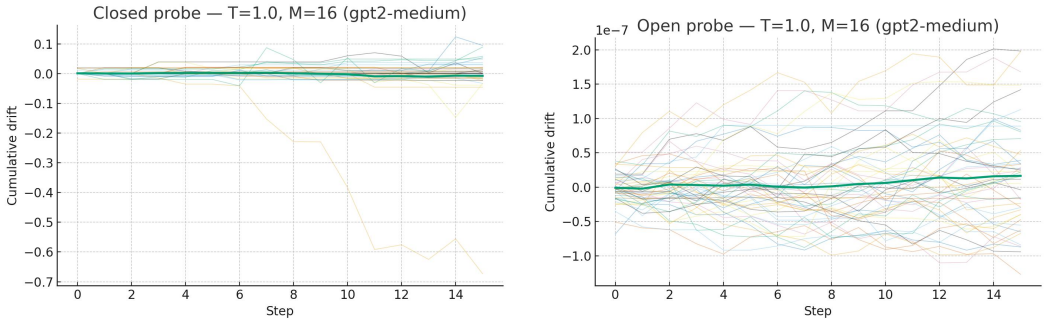
Table 3: Ablation neutrality audits for `gpt2-medium` under varying temperature T and sibling count M. Closed probes show centered but wandering cumulative drift; open probes remain numerically tiny by comparison. $E_{\max}$ values near 1 indicate no evidence of sustained bias; for context, an anytime e-test would only approach rejection around $E_{\max} \gtrsim 20$ at $\alpha = 0.05$.

Model	Block	$\hat{\mu}$	SE	95% CI
`tiny-gpt2`	All (4)	$1.157e{-07}$	$9.729e{-08}$	$(-1.877e{-10},\ 3.468e{-07})$
`distilgpt2`	All (4)	$-2.833e{-04}$	$9.159e{-03}$	$(-1.839e{-02},\ 1.843e{-02})$
`gpt2-medium`	All (4)	$-2.844e{-04}$	$8.178e{-03}$	$(-1.687e{-02},\ 1.630e{-02})$
`gpt2-large`	All (4)	$1.221e{-04}$	$1.341e{-02}$	$(-2.851e{-02},\ 2.788e{-02})$

Table 4: Layer-as-agent diagnostics aggregated across residual blocks. Reported are the mean action $\hat{\mu}$, its standard error, and a 95% confidence interval. Intervals cover zero throughout, indicating no systematic bias at the block level.

F.1 Layer perspective.

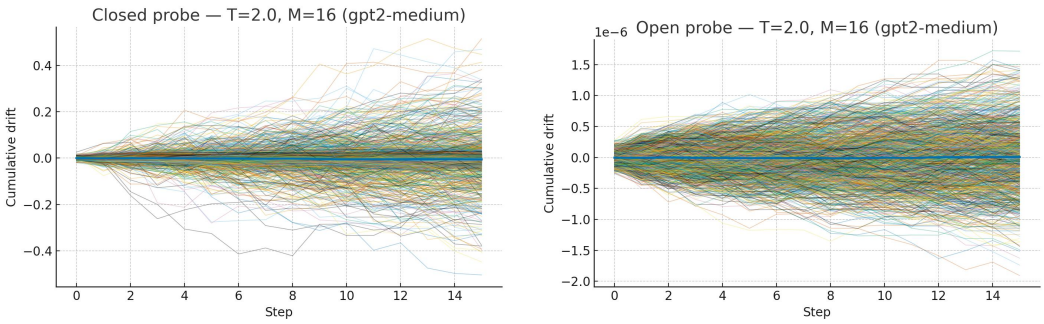
F.2 Trajectory-level Neutrality Audits with respect to T



Closed probe. Open probe.

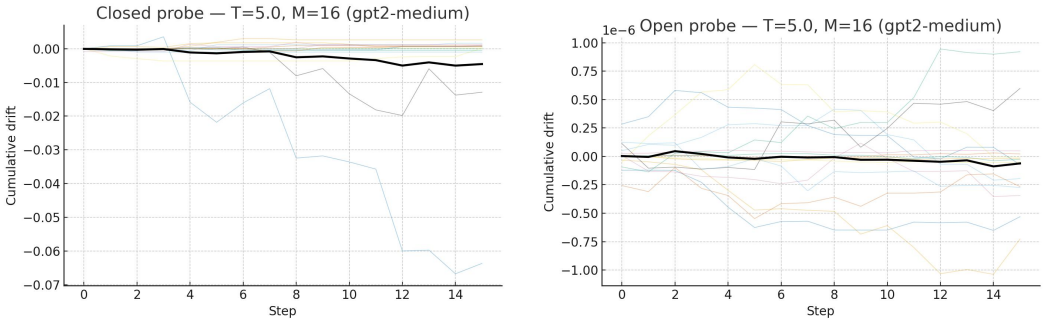
Figure 3: Neutrality audit for `gpt2-medium` with $T = 1, M = 16$.

Remark regarding $T = 5$. At high temperature the softmax flattens, increasing token entropy and branching variance in the open probe; closed increments remain martingale differences, but their step variance also grows because re-embeddings explore more of the state space. In Table 3 (T=5, M=16) the prompt-level t-test is marginal ($p = 3.20\times10^{-2}$) around a very small mean drift (6.19×10^{-5}), yet the anytime e-test stays near one ($E_{\max} = 1.005$), far below rejection thresholds (e.g., ≥ 20 at α=0.05), indicating no sustained deviation. Trajectories therefore look more volatile (variance inflation) but remain neutral in expectation. Moreover, theory predicts a smaller corridor at higher T (softmax Lipschitz $1/(2T)$), consistent with the absence of bias despite noisier paths.



Closed probe. Open probe.

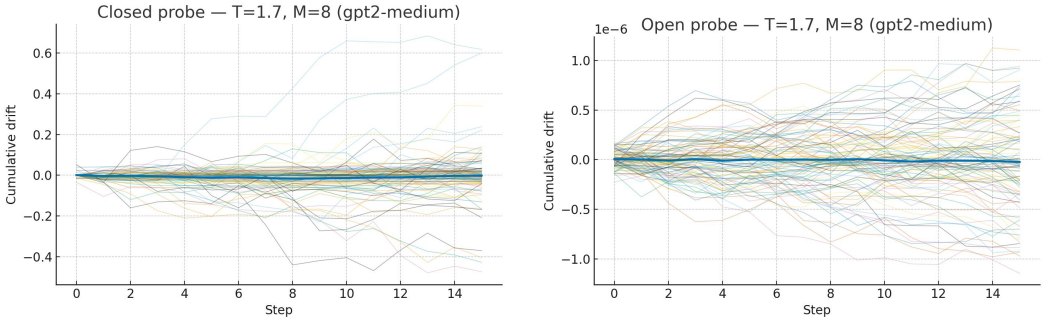
Figure 4: Neutrality audit for `gpt2-medium with` $T = 2, M = 16$.



Closed probe. Open probe.

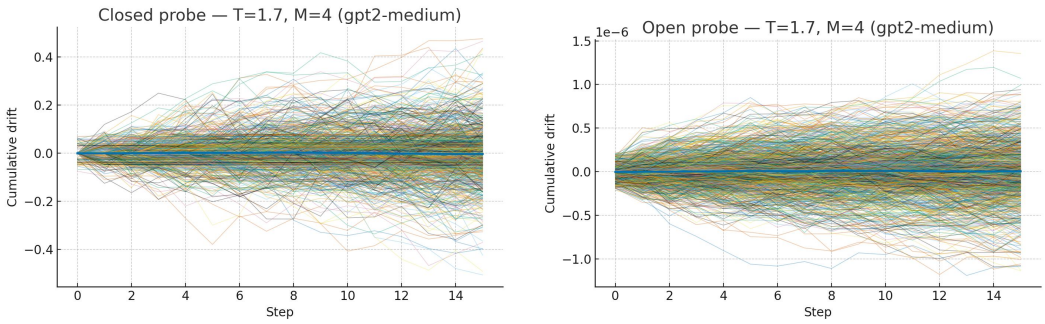
Figure 5: Neutrality audit for `gpt2-medium with` $T = 5, M = 16$.

F.3 Trajectory-level Neutrality Audits with respect to M



Closed probe. Open probe.

Figure 6: Neutrality audit for `gpt2-medium with` $T = 1.7, M = 8$.



Closed probe. Open probe.

Figure 7: Neutrality audit for `gpt2-medium with` $T = 1.7, M = 4$.