ORIGINAL RESEARCH ARTICLE | PREPRINT FOR *NATURE MACHINE INTELLIGENCE* / arXiv

The Contextuality–Signaling Dichotomy in Autoregressive Large Language Models: Causal-Mask Absorption, Semantic Frustration Phase Transition, and Correlated Effective Multiplicity

Fabien Pierret

Independent Researcher | ORCID: 0009-0001-1223-2170

Full Reproducibility: All empirical results reported in this manuscript come from deterministic live evaluations (`seed = 42`, exact next-token logits, zero Monte Carlo error on cyclic distributions) executed on `mlx-community/Qwen2.5-1.5B-Instruct-4bit` (Apple Silicon M4 Max, `MLX 0.32.4`), with cryptographically verified invariance of all 732 frozen weight tensors (`SHA-256 = e2d0d0869ca0b89e7f4106fd1e49e0b5d37e39876e83de15cb39e0edff74c573`, $\|\Delta\theta^*\|_2 = 0.0$), via the accompanying scripts `exp8_real_llm_prompt_dredging.py`, `exp9_real_llm_cbd_signaling.py`, and `exp11_paper2_autoregressive_frustration_transition.py`.

Abstract

Recent empirical studies have detected tens of millions of *Contextuality-by-Default* (CbD) contextual instances in bidirectional masked language models such as BERT [1], fueling debate over whether modern Transformers exhibit quantum-like nonclassicality or merely sequential classical priming. In this work, we resolve this ambiguity for **autoregressive, causally masked Large Language Models (LLMs)** by proving three exact mathematical theorems and validating them live on frozen `Qwen2.5-1.5B-Instruct-4bit` weights ($\|\Delta\theta^*\|_2 = 0$).

First, we establish the **Causal-Mask Signaling-Absorption Theorem (Theorem 1)**: decomposing any directed autoregressive n-cycle into Turn-1 causal marginals m_i, answer-independent prompt shifts $\bar{\mu}_i$, and answer-dependent KV-cache transition contrasts δ_i, we derive an exact identity showing that pure prompt-framing shifts ($\delta_i = 0$), high-confidence Turn-1 saturation ($1 - m_i^2 \to 0$), and signaling-dominated edges ($|\Delta m_{i+1}| \geq (1 + |m_i|)|\delta_i|$) force the KV-cache marginal inconsistency Δ_0 to absorb **100%** of raw Bell–CHSH-like excesses ($C_{\mathrm{CbD}} \equiv s_{\mathrm{odd}} - \Delta_0 \leq n - 2$), even when s_{odd} approaches the algebraic ceiling n. This explains why **38 / 51 (74.5%)** natural multi-turn cycles on `Qwen2.5-1.5B` (Exp. 9 + Exp. 11 Tier 2) breach $s_{\mathrm{odd}} > 2$ (up to 3.487) while **0 / 51 (0.0%)** violate CbD ($\max C_{\mathrm{CbD}} = 1.9997$).

Second, we prove the **Necessary and Sufficient Autoregressive Frustration Transition Criterion (Theorem 2)**, identifying the exact algebraic phase boundary where an autoregressive LLM crosses $C_{\mathrm{CbD}} > n - 2$ (requiring simultaneous Turn-1 uncertainty $1 - m_i^2 > 0$, answer-dependent coupling $\delta_i \neq 0$, and odd-parity frustration $\overline{|\delta|} > 1 - 2/n$ in balanced regimes), together with a closed-form **Möbius Symmetrization Law** proving that 50/50 forward–reverse symmetrization reduces ideal frustrated signaling to $\Delta_0^{\mathrm{sym}} = \max(|m_1|, |m_3|) + \max(|m_2|, |m_4|) \leq 2$, forcing $C_{\mathrm{CbD}}^{\mathrm{sym}} > 2$ for 100% of unsaturated cycles. Live experiments on `Qwen2.5-1.5B` (Exp. 11 Part A) confirm this phase transition directly: balanced Möbius cycle `M01` ($\overline{|m_1|} = 0.227$) reaches $C_{\mathrm{CbD}} = 2.959$ (forward), 3.043 (reverse), and 3.209 (symmetrized)—surpassing even Tsirelson's quantum bound $2\sqrt{2} \approx 2.828$ via purely classical KV-cache joint access—while polarized Möbius cycles remain absorbed below 2 in directed mode yet cross $C_{\mathrm{CbD}}^{\mathrm{sym}} > 2$ in **15 / 15 (100%)** symmetrized evaluations. We rigorously separate this semantic classical contextuality (faithfully lifted to bond-dimension $\chi \leq 2$ Matrix Product States) from physical Bell nonlocality (0/1,050 violations under isolated $\eta = 1$ queries, $\max|S| = 1.9999$), and demonstrate live on `Qwen2.5-1.5B` (Exp. 11 Part B) that even benign confidence-threshold abstention ($\eta \leq 95\%$) opens a real-LLM **detection loophole** ($|S| = 2.2756$; guardrail $|S| = 2.3537$), immediately closed by deterministic $\emptyset \mapsto +1$ imputation ($|S| \leq 1.7783$).

Third, we prove the **Correlated Effective Prompt Multiplicity Theorem (Theorem 3)** for post-selective *prompt dredging*, establishing exact Slepian bounds, a universal spectral ordering $1 \leq M_{\mathrm{eff}}^{\mathrm{PR}} \leq M_{\mathrm{eff}}^{\mathrm{Galwey}} \leq M$ and $M_{\mathrm{eff}}^{\mathrm{PR}} \leq M_{\mathrm{eff}}^{\mathrm{NC}} \leq M$, and an active-set variance ceiling that jointly explain both the 10×

multiplicity divergence observed in Exp. 8 ($M_{\text{eff}} = 17.2$ vs. 175.5) and the 2.6×–2.8× error reduction achieved by spectral estimators on live `Qwen2.5-1.5B` prompt pools (Exp. 11 Part C).

Keywords: Contextuality-by-Default (CbD); Causal attention mask; Autoregressive Large Language Models; KV-cache signaling; Parity-frustration phase transition; Detection loophole; Prompt dredging; Spectral effective multiplicity; Matrix Product States (MPS); Kirkwood–Dirac quasiprobabilities.

1. Introduction and Scientific Problem Statement

The emergence of seemingly non-commutative behaviors in Large Language Models (LLMs)—order effects in multi-turn dialogues, extreme sensitivity to few-shot example ordering [2], and framing biases reminiscent of the human survey order effects studied by Wang et al. [3]—has revived interest in applying formalisms from **quantum contextuality** directly to Transformer architectures. In December 2024, Lo et al. [1] (`arXiv:2412.16806`) reached a striking empirical milestone by identifying 36,938,948 **contextual instances** under the *Contextuality-by-Default* (CbD) framework [4–6] inside **BERT**, a bidirectional masked language model with unmasked attention.

This finding immediately opens **three fundamental questions** that have remained analytically unresolved in the literature:

1. **The causal-mask enigma in autoregressive LLMs:** Unlike BERT—where two tokens inside the same context attend to each other symmetrically within a bidirectional window—modern autoregressive LLMs (the GPT, Llama, and Qwen families) enforce a **triangular causal attention mask**: the Turn-1 answer (Q_i) is sampled *strictly before* the Turn-2 question (q_{i+1}) enters the context, whereas Turn 2 attends to the entire history (q_i, a) through the **Key–Value (KV) cache**. In our initial live measurements on `Qwen2.5-1.5B-Instruct-4bit` (Experiment 9), a stark paradox emerged: **27 / 36 (75%)** of standard multi-turn cycles strongly violate the naive Bell–CHSH inequality [7, 8] ($s_{\text{odd}} > 2$, peaking at 3.4872—well above Tsirelson's quantum bound $2\sqrt{2} \approx 2.8284$ [9]), **yet exactly 0 / 36 (0.0%)** cross the CbD noncontextuality threshold once the KV-cache marginal signaling Δ_0 is subtracted ($\max C_{\text{CbD}} = 1.9997$). Is this absolute zero a universal structural law of autoregressive LLMs, or does there exist a **constructive phase transition** enabling an autoregressive model to genuinely violate $C_{\text{CbD}} > 2$?

2. **Epistemological demarcation against *quantum woo* and the detection-loophole trap:** If an autoregressive LLM can cross $C_{\text{CbD}} > 2$ (and even exceed Tsirelson's bound $2\sqrt{2}$), how can we prove unambiguously that this phenomenon is **classical semantic contextuality mediated by joint KV-cache access**, strictly distinct from physical Bell nonlocality ($|S_{\text{CHSH}}| \leq 2$ under causal isolation), and how can we prevent abstention or refusal filters (*rejection sampling*, confidence thresholds, safety guardrails) from opening an artificial **detection loophole** in LLM evaluation benchmarks?

3. **Spectral geometry of post-selection over correlated prompts (*prompt dredging*):** When selecting the highest-scoring prompt p^* from M candidate formulations on frozen model weights θ^*, prompt predictions are heavily correlated ($\bar{\rho} = 0.9564$ under null labels vs. $\bar{\rho} = 0.2490$ on real sentiment classification in Experiment 8). What exact mathematical law connects the eigenvalue spectrum of the prompt correlation matrix Σ_{prompt} and signal heterogeneity to the optimism gap $\mathbb{E}[\Delta_{\text{IS}}(p^*) - \Delta_{\text{OOS}}(p^*)]$?

This article provides a **unified, analytically exact, and empirically verified answer on live `Qwen2.5-1.5B` weights** to all three questions.

2. Systematic Prior-Art Audit (2014–2026)

In accordance with high-rigor publication standards (*Nature Machine Intelligence* / *PRX Quantum* / *NeurIPS*), Table 1 demarcates established results in the verified 2014–2026 literature from the **original theoretical and empirical contributions** of this work.

Research Axis	Established Prior Art (2014–2026)	Unresolved Gap	Novel Contribution of This Paper
CbD contextuality and cyclic systems with signaling	Kujala et al. [4] and Dzhafarov and Kujala [5], Dzhafarov et al. [6] establish the exact criterion $s_{\text{odd}} - \Delta_0 \leq n - 2$ for arbitrary binary n-cycles. Wang et al. [3] establish the QQ equality ($q_{\text{QQ}} = 0$) in human judgment.	No algebraic decomposition tailored to the asymmetric factorization of an **autoregressive causal mask** $P(a \mid q_i)P(b \mid q_i, a, q_{i+1})$ in terms of $(m_i, \bar{\mu}_i, \delta_i)$.	**Theorem 1 (*Causal-Mask Absorption*):** exact master identity (5), total absorption at $\delta_i = 0$ (comonotonic coupling), Turn-1 variance envelope, edgewise criterion $\lvert\Delta m_{i+1}\rvert \geq (1 + \lvert m_i\rvert)\lvert\delta_i\rvert$, and sharpness counterexample (2.0051 rounded / 2.0052 unrounded).
Contextuality in Large Language Models	Lo et al. [1] (`arXiv:2412.16806`) detect 36.9M CbD instances in **BERT** (symmetric bidirectional attention on noun–adjective pairs).	No theoretical or experimental characterization of the **dichotomy between natural cycles (0% CbD) and constructive parity-frustration transitions** in causally masked autoregressive LLMs.	**Theorem 2 (Exact IFF transition criterion + Möbius Symmetrization Law** $\Delta_0^{\mathbf{sym}} = \max(\lvert m_1\rvert, \lvert m_3\rvert) + \max(\lvert m_2\rvert, \lvert m_4\rvert) \leq 2$**)** and live verification on `Qwen2.5-1.5B` (0/51 natural vs. `M01` at 2.959/3.043/3.209 and 15/15 symmetrized).
MPS lifting, Kirkwood–Dirac negativity, and Detection Loophole	Bradley et al. [10] lift classical sequences to MPS; Lostaglio et al. [11] link Kirkwood–Dirac negativity to contextuality; Pearle [12], Garg and Mermin [13], Eberhard [14], and Larsson [15] characterize optical detection loopholes.	Never unified to separate KV-cache semantic contextuality ($\chi \leq 2$), isolated Bell nonlocality ($\lvert S\rvert \leq 2$), and **confidence/refusal detection loopholes on real LLM logits**.	**Proposition 1 + Exp. 11 Part B:** first live empirical demonstration on real LLM logits that benign confidence abstention alone ($\eta \leq 95\%$) breaches $\lvert S\rvert \leq 2$ (2.2756) and that deterministic $\emptyset \mapsto +1$ imputation closes the loophole (≤ 1.7783).
Correlated effective multiplicity in *prompt dredging*	Cheverud [16], Nyholt [17], and Galwey [18] introduce M_{eff} in genetics; Bailey and López de Prado [19] formalize multiple-testing selection bias; Slepian [20] and Boucheron et al. [21] bound correlated Gaussian maxima; Lu et al. [2] study few-shot order sensitivity.	No theorem unifying the **exact spectral orderings** ($M_{\text{eff}}^{\text{PR}}$, $M_{\text{eff}}^{\text{Galwey}}$, $M_{\text{eff}}^{\text{NC}}$), **PR asymptotic saturation**, the **variance ceiling** σ_k^2, and live validation on frozen LLM weights.	**Theorem 3 + Proposition 2 + Exp. 8 & 11 Part C:** exact proofs of $1 \leq M_{\text{eff}}^{\text{PR}} \leq M_{\text{eff}}^{\text{Galwey}} \leq M$ and $M_{\text{eff}}^{\text{PR}} \leq M_{\text{eff}}^{\text{NC}} \leq M$, analytical PR-saturation vs. Galwey-envelope diagnosis, and 2.6×–2.8× MAE reduction live on `Qwen2.5-1.5B`.

Table 1: Systematic prior-art audit verified across the 2014–2026 literature, strictly separating established foundations from the novel contributions of this manuscript.

3. Formal Framework: Causally Masked Autoregressive Cyclic Systems

Let an autoregressive language model with frozen parameters θ^* define, for any text prefix h, an exact next-token probability distribution $P_{\theta^*}(\cdot \mid h)$ over the vocabulary. For a set of $n \geq 3$ binary questions $(q_1, \ldots, q_n)$ (indices taken modulo n, so $q_{n+1} \equiv q_1$), restricted to normalized binary

answers $a, b \in \{-1, +1\}$ (e.g., No $\leftrightarrow -1$, Yes $\leftrightarrow +1$), we evaluate the n **directed sequential contexts** $c_i = (q_i \to q_{i+1})$ for $i \in \mathbb{Z}_n$.

By virtue of the **causal attention mask** (which prohibits the Turn-1 token position from attending to future Turn-2 tokens), the exact joint probability distribution in context c_i factorizes asymmetrically:

$$P_{c_i}(Q_i = a,\, Q_{i+1} = b) \;=\; P(Q_i = a \mid q_i)\, P(Q_{i+1} = b \mid q_i, a, q_{i+1}), \qquad a, b \in \{-1, +1\}. \tag{1}$$

3.1. Complete Canonical Parametrization of an Autoregressive n-Cycle

Every directed edge $i \in \mathbb{Z}_n$ of the form (1) stands in exact bijection with a real parameter triple $(m_i, \bar{\mu}_i, \delta_i) \in [-1, 1]^3$ satisfying the feasibility constraint $|\bar{\mu}_i| + |\delta_i| \le 1$:

1. **Turn-1 causal marginal** (strictly independent of q_{i+1} due to the causal mask):
$$m_i \;\equiv\; \mathbb{E}_{c_i}[Q_i] \;=\; 2P(Q_i = +1 \mid q_i) - 1 \;\in\; [-1, +1].$$
2. **Turn-2 conditional expectations** given the realized Turn-1 answer $a \in \{+1, -1\}$:
$$\mu_i^+ \;\equiv\; \mathbb{E}[Q_{i+1} \mid q_i, +1, q_{i+1}], \qquad \mu_i^- \;\equiv\; \mathbb{E}[Q_{i+1} \mid q_i, -1, q_{i+1}] \;\in\; [-1, +1].$$
3. **Orthogonal decomposition of KV-cache transfer** into an **answer-independent contextual shift** ($\bar{\mu}_i$) and an **answer-dependent transition contrast** (δ_i):
$$\begin{aligned} \bar{\mu}_i &\equiv \frac{\mu_i^+ + \mu_i^-}{2}, \\ \delta_i &\equiv \frac{\mu_i^+ - \mu_i^-}{2} = P(Q_{i+1} = +1 \mid q_i, +1, q_{i+1}) - P(Q_{i+1} = +1 \mid q_i, -1, q_{i+1}). \end{aligned} \tag{2}$$

Lemma 1 (Exact Turn-2 Marginal and Edge Correlator Identities). *For any autoregressive edge (1) parametrized by $(m_i, \bar{\mu}_i, \delta_i)$, the Turn-2 marginal $\tilde{m}_{i+1} \equiv \mathbb{E}_{c_i}[Q_{i+1}]$ and the edge correlator $E_i \equiv \mathbb{E}_{c_i}[Q_i Q_{i+1}]$ satisfy identically:*

$$\tilde{m}_{i+1} \;=\; \bar{\mu}_i + m_i\, \delta_i, \qquad E_i \;=\; \delta_i + m_i\, \bar{\mu}_i \;=\; m_i\, \tilde{m}_{i+1} + (1 - m_i^2)\, \delta_i. \tag{3}$$

Proof. Using $P(Q_i = \pm 1 \mid q_i) = \frac{1 \pm m_i}{2}$ and $\mu_i^\pm = \bar{\mu}_i \pm \delta_i$:

$$\begin{aligned} \tilde{m}_{i+1} &= \frac{1 + m_i}{2}(\bar{\mu}_i + \delta_i) + \frac{1 - m_i}{2}(\bar{\mu}_i - \delta_i) = \bar{\mu}_i + m_i \delta_i, \\ E_i &= \frac{1 + m_i}{2}(\bar{\mu}_i + \delta_i) - \frac{1 - m_i}{2}(\bar{\mu}_i - \delta_i) = \delta_i + m_i \bar{\mu}_i. \end{aligned}$$

Substituting $\bar{\mu}_i = \tilde{m}_{i+1} - m_i \delta_i$ into E_i yields $E_i = \delta_i + m_i(\tilde{m}_{i+1} - m_i \delta_i) = m_i \tilde{m}_{i+1} + (1 - m_i^2)\delta_i$ (which expresses the exact covariance identity $\mathrm{Cov}_{c_i}(Q_i, Q_{i+1}) = \mathrm{Var}(Q_i)\, \delta_i$). □

3.2. Exact *Contextuality-by-Default* Criterion

Let $\Delta m_{i+1} \equiv \tilde{m}_{i+1} - m_{i+1}$ denote the marginal shift undergone by question q_{i+1} when asked at Turn 2 after q_i rather than at Turn 1 in isolation. The **total marginal signaling** of the cycle and the **odd-parity correlation functional** are defined as:

$$\Delta_0 \;\equiv\; \sum_{i=1}^{n} \left| \mathbb{E}_{c_i}[Q_i] - \mathbb{E}_{c_{i-1}}[Q_i] \right| \;=\; \sum_{i=1}^{n} |\tilde{m}_{i+1} - m_{i+1}|, \qquad s_{\mathrm{odd}} \;\equiv\; \max_{\boldsymbol{\iota} \in \mathcal{I}_n^{\mathrm{odd}}} \sum_{i=1}^{n} \iota_i E_i,$$

where $\mathcal{I}_n^{\mathrm{odd}} \equiv \{\boldsymbol{\iota} \in \{-1, +1\}^n : \prod_{i=1}^n \iota_i = -1\}$. By Kujala et al. [4], the binary n-cycle is **CbD-noncontextual** (i.e., admits a maximal coupling over all $2n$ contextual random variables achieving $\Pr(Q_i^{c_i} = Q_i^{c_{i-1}}) = 1 - \frac{1}{2}|m_i - \tilde{m}_i|$ for every i) **if and only if**:

$$C_{\mathrm{CbD}} \;\equiv\; s_{\mathrm{odd}} - \Delta_0 \;\le\; n - 2. \tag{4}$$

For $n = 4$, the noncontextuality ceiling is therefore $C_{\mathrm{CbD}} \le 2$. Furthermore, for any ordered pair (A, B), the **QQ-equality** residual of Wang et al. [3] is $q_{\mathrm{QQ}}(A, B) \equiv [P(1,1) + P(-1,-1)]_{A \to B} - [P(1,1) + P(-1,-1)]_{B \to A}$, which vanishes identically under projective quantum Lüders measurements without prior signaling, but generically diverges under asymmetric KV-cache signaling.

4. Theorem 1: Causal-Mask Signaling-Absorption

We now state and prove our first main theorem, which exposes the exact algebraic mechanism by which KV-cache signaling Δ_0 absorbs raw correlation excesses $s_{\text{odd}} - (n-2)$ in autoregressive models.

Theorem 1 (Causal-Mask Signaling-Absorption Theorem). *Let an autoregressive binary n-cycle ($n \geq 3$) have canonical parameters $(m_i, \bar{\mu}_i, \delta_i)_{i=1}^n$.*

1. ***(Master Causal-Absorption Identity)*** *For every odd-parity sign vector $\boldsymbol{\iota} \in \mathcal{I}_n^{odd}$, writing $\Delta m_{i+1} \equiv \tilde{m}_{i+1} - m_{i+1}$:*

$$\sum_{i=1}^{n} \iota_i E_i - \Delta_0 \;=\; (n-2) \;-\; \mathcal{G}(\boldsymbol{\iota}, \mathbf{m}) \;-\; \mathcal{A}(\boldsymbol{\iota}, \mathbf{m}, \tilde{\mathbf{m}}) \;+\; \sum_{i=1}^{n} \iota_i \,(1 - m_i^2)\,\delta_i, \tag{5}$$

   where:

   - *the **marginal polytope deficit** $\mathcal{G}(\boldsymbol{\iota}, \mathbf{m}) \equiv (n-2) - \sum_{i=1}^n \iota_i m_i m_{i+1}$ always satisfies $\mathcal{G}(\boldsymbol{\iota}, \mathbf{m}) \geq 0$;*
   - *the **signaling-absorption term** $\mathcal{A}(\boldsymbol{\iota}, \mathbf{m}, \tilde{\mathbf{m}}) \equiv \sum_{i=1}^n \left(|\Delta m_{i+1}| - \iota_i m_i \Delta m_{i+1}\right)$ always satisfies $\mathcal{A}(\boldsymbol{\iota}, \mathbf{m}, \tilde{\mathbf{m}}) \geq \sum_{i=1}^n (1 - |m_i|)\,|\Delta m_{i+1}| \geq 0$.*

2. ***(Corollary 1A — Total Absorption under Pure Prompt-Framing Shifts $\delta_i = 0$)*** *If the answer-dependent transition contrast vanishes on every edge ($\delta_i = 0\ \forall i \in \mathbb{Z}_n$), then for **all** arbitrary contextual shifts $\bar{\mu}_i \in [-1, +1]$ and **all** Turn-1 marginals $m_i \in [-1, +1]$—even when the raw functional s_{odd} approaches the algebraic ceiling n—one has identically:*

$$C_{CbD} \;=\; s_{odd} - \Delta_0 \;\leq\; n - 2. \tag{6}$$

3. ***(Corollary 1B — Turn-1 Variance Envelope and Edgewise Signaling Dominance)*** *For any autoregressive transition matrix (even maximally frustrated $|\delta_i| = 1$):*

$$C_{CbD} \;\leq\; (n-2) + \sum_{i=1}^{n} (1 - m_i^2)\,|\delta_i|. \tag{7}$$

   In particular, whenever marginal signaling dominates answer-dependent coupling on every edge in the sense that

$$|\tilde{m}_{i+1} - m_{i+1}| \;\geq\; (1 + |m_i|)\,|\delta_i| \qquad \forall i \in \mathbb{Z}_n, \tag{8}$$

   *absorption is **complete**: $C_{CbD} \leq n - 2$ identically.*

4. ***(Corollary 1C — Balanced Turn-1 Regime $\mathbf{m} = \mathbf{0}$)*** *When all Turn-1 marginals are balanced ($m_i = 0\ \forall i$), identity (5) simplifies exactly to:*

$$C_{CbD}\big|_{\mathbf{m}=\mathbf{0}} \;=\; s_{odd}(\boldsymbol{\delta}) - \sum_{i=1}^{n} |\bar{\mu}_i|. \tag{9}$$

   In particular, for any homogeneous unfrustrated coupling ($\delta_1 = \cdots = \delta_n = \delta \geq 0$), $s_{odd}(\boldsymbol{\delta}) = (n-2)\delta \leq n - 2$, hence $C_{CbD}|_{\mathbf{m}=\mathbf{0}} \leq n - 2$.

Proof. **Step 1 (Master Identity (5)).** By Lemma 1 (Eq. (3)), $E_i = m_i \tilde{m}_{i+1} + (1 - m_i^2)\delta_i$. Adding and subtracting $m_i m_{i+1}$ gives:

$$E_i = m_i m_{i+1} + m_i(\tilde{m}_{i+1} - m_{i+1}) + (1 - m_i^2)\delta_i = m_i m_{i+1} + m_i \Delta m_{i+1} + (1 - m_i^2)\delta_i.$$

Summing with signs ι_i and subtracting $\Delta_0 = \sum_{i=1}^n |\Delta m_{i+1}|$ yields (5) identically.

Step 2 (Nonnegativity of $\mathcal{G}$ and $\mathcal{A}$).

- The function $\Phi(\mathbf{m}) = \sum_{i=1}^{n} \iota_i m_i m_{i+1}$ is multilinear in each coordinate $m_k \in [-1, +1]$ and therefore attains its maximum at a vertex of the hypercube $\{-1, +1\}^n$. At any vertex $m_i \in \{-1, +1\}$, the product of the n summands equals $\prod_{i=1}^{n}(\iota_i m_i m_{i+1}) = (\prod_{i=1}^{n} \iota_i)\,(\prod_{i=1}^{n} m_i^2) = (-1) \cdot 1 = -1$. Hence at least one of the n terms in $\{-1, +1\}$ must equal -1, forcing $\Phi(\mathbf{m}) \leq (n-1) - 1 = n - 2$ and thus $\mathcal{G}(\boldsymbol{\iota}, \mathbf{m}) \geq 0$.
- On every edge i, $|\iota_i m_i \Delta m_{i+1}| = |m_i|\,|\Delta m_{i+1}|$, which immediately gives:

$$|\Delta m_{i+1}| - \iota_i m_i \Delta m_{i+1} \;\geq\; (1 - |m_i|)\,|\Delta m_{i+1}| \;\geq\; 0.$$

Step 3 (Corollary 1A — Dual Algebraic and Constructive Maximal-Coupling Proof). If $\delta_i = 0$ for all i, the final sum in (5) vanishes identically, so $\sum_i \iota_i E_i - \Delta_0 = (n-2) - \mathcal{G} - \mathcal{A} \leq n - 2$ for every $\boldsymbol{\iota} \in \mathcal{I}_n^{\text{odd}}$, giving $C_{\text{CbD}} \leq n - 2$. (*Independent constructive probabilistic proof:* let $U_1, \ldots, U_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, 1]$ and set $p_i = \frac{1+m_i}{2}$, $\tilde{p}_{i+1} = \frac{1+\bar{\mu}_i}{2}$. The variables $Q_i^{c_i} \equiv \text{sgn}(p_i - U_i)$ and $Q_{i+1}^{c_i} \equiv \text{sgn}(\tilde{p}_{i+1} - U_{i+1})$ are independent within each context c_i—reproducing P_{c_i} at $\delta_i = 0$ exactly—while sharing the same uniform U_i across c_i and c_{i-1}, achieving $\Pr(Q_i^{c_i} \neq Q_i^{c_{i-1}}) = |p_i - \tilde{p}_i| = \frac{1}{2}|m_i - \tilde{m}_i|$, an exact maximal coupling.) Note that s_{odd} may nevertheless reach n: for $n = 4$, choosing $\mathbf{m} = (1, 1, 1, 1)$ and $\bar{\boldsymbol{\mu}} = (1, 1, 1, -1)$ gives $\mathbf{E} = (1, 1, 1, -1) \implies s_{\text{odd}} = 4$, yet $\Delta_0 = |-1 - 1| = 2 \implies C_{\text{CbD}} = 4 - 2 = 2$.

Step 4 (Corollary 1B). From (5), $\mathcal{G} \geq 0$ and $\mathcal{A} \geq 0$ immediately imply (7). Using the exact factorization $1 - m_i^2 = (1 - |m_i|)(1 + |m_i|)$ together with $\mathcal{A}(\boldsymbol{\iota}) \geq \sum_i (1 - |m_i|)|\Delta m_{i+1}|$:

$$\sum_{i=1}^{n} \iota_i (1 - m_i^2)\delta_i - \mathcal{A}(\boldsymbol{\iota}) \;\leq\; \sum_{i=1}^{n} (1 - |m_i|)\Big[(1 + |m_i|)|\delta_i| - |\Delta m_{i+1}|\Big].$$

Under condition (8), every bracketed term is ≤ 0, so $C_{\text{CbD}} \leq n - 2$ identically.

Step 5 (Corollary 1C). At $m_i = 0$, $E_i = \delta_i$, $\tilde{m}_{i+1} = \bar{\mu}_i$, and $\Delta_0 = \sum_{i=1}^{n} |\bar{\mu}_i|$, yielding (9) directly. □

Sharpness Remark *(Adversarial Global-Optimization Audit).* It is tempting—yet **mathematically false**—to conjecture that every homogeneous monotone coupling $\delta_1 = \cdots = \delta_n = \delta \geq 0$ remains noncontextual when $\mathbf{m} \neq \mathbf{0}$. Adversarial multi-start Powell optimization uncovers an explicit $n = 4$ counterexample at $\delta = 0.8871$, $\mathbf{m} = (0.7691, 0.7870, 0.8304, 0.7398)$, and $\bar{\boldsymbol{\mu}} = (0.1048, 0.1129, 0.0031, 0.1129)$, which achieves $C_{\mathbf{CbD}} = 2.0051 > 2$ by direct evaluation on these 4-decimal rounded parameters ($C_{\text{CbD}} = 2.0052$ at the unrounded global optimum; for $n = 3$ at $\delta = 0.4132$, $C_{\text{CbD}} = 1.0361 > 1$). Likewise, a non-homogeneous even-sign cycle at $\mathbf{m} = \mathbf{0}$ and $\boldsymbol{\delta} = (1, 1, 1, 0)$ attains $C_{\text{CbD}} = 3 > 2$. Master identity (5) and Corollaries 1A–1C are therefore **exact and sharp**: none of the sufficient conditions $\delta_i = 0$ (1A), signaling dominance (8), or balanced marginals $\mathbf{m} = \mathbf{0}$ (1C) can be arbitrarily relaxed without exposing the contextuality boundary.

5. Theorem 2: Exact Phase Transition Criterion, Möbius Symmetrization Law, and MPS / Kirkwood–Dirac Lifting

Master identity (5) is not merely an upper bound: it provides the **exact necessary and sufficient condition** for semantic contextuality to emerge in an autoregressive LLM, together with the closed-form law governing the 100% symmetrized transition observed in Experiment 11.

Theorem 2 (Autoregressive Phase Transition to Semantic Contextuality). *Let an autoregressive n-cycle ($n \geq 3$) have parameters $(m_i, \bar{\mu}_i, \delta_i)_{i=1}^{n}$ satisfying $|\bar{\mu}_i| + |\delta_i| \leq 1$.*

1. ***(Exact Necessary and Sufficient Criterion)*** *The system is CbD-contextual ($C_{CbD} > n-2$) **if and only if** there exists an odd-parity sign vector $\boldsymbol{\iota} \in \mathcal{I}_n^{odd}$ such that:*

$$\sum_{i=1}^{n} \iota_i\,(1 - m_i^2)\,\delta_i \;>\; \mathcal{G}(\boldsymbol{\iota}, \mathbf{m}) \;+\; \mathcal{A}(\boldsymbol{\iota}, \mathbf{m}, \tilde{\mathbf{m}}). \tag{10}$$

Consequently, the simultaneous presence of ***two ingredients*** *is strictly necessary: (i) nonzero Turn-1 uncertainty* ($\exists i : |m_i| < 1$) *and (ii) answer-dependent KV-cache coupling* ($\exists i : \delta_i \neq 0$).

2. ***(Constructive Odd-Parity Frustration Threshold)*** *Suppose the transitions implement* ***odd-parity frustration*** $\prod_{i=1}^{n} \mathrm{sgn}(\delta_i) = -1$ *(for* $n = 4$*: three agreement edges* $\delta_1, \delta_2, \delta_3 > 0$ *and one inversion edge* $\delta_4 < 0$*, a* ***Möbius cycle*** *topology), with mean contrast magnitude* $\overline{|\delta|} \equiv \frac{1}{n}\sum_{i=1}^{n} |\delta_i|$.

   - *In the balanced regime* ($m_i = \bar{\mu}_i = 0\ \forall i$)*, the cycle is CbD-contextual* ***if and only if****:*

$$\overline{|\delta|} \;>\; 1 - \frac{2}{n} \qquad \left(i.e.,\ \overline{|\delta|} > \frac{1}{2} \ \text{ for } n = 4 \right), \tag{11}$$

     with $C_{CbD} = n\,\overline{|\delta|}$*, reaching the* ***algebraic ceiling*** n *(i.e.,* $4 > 2\sqrt{2}$ *for* $n = 4$*) as* $|\delta_i| \to 1$.

   - *Under bounded Turn-1 polarization* $|m_i| \leq \varepsilon$ *and contextual midpoint bias* $|\bar{\mu}_i| \leq \eta$*, the CbD witness obeys the guaranteed constructive lower bound:*

$$C_{CbD} \;\geq\; n\,\overline{|\delta|}(1-\varepsilon) \;-\; n\,\varepsilon \;-\; n(1+\varepsilon)\,\eta. \tag{12}$$

3. ***(Exact Möbius Symmetrization Law for*** $n = 4$***)*** *Consider an ideal 4-cycle Möbius system* ($\delta_1 = \delta_2 = \delta_3 = +1$, $\delta_4 = -1$, $\bar{\mu}_i = 0$ *in both forward* $\rightarrow$ *and reverse* $\leftarrow$ *directions) with arbitrary Turn-1 marginals* $\mathbf{m} = (m_1, m_2, m_3, m_4) \in (-1, +1)^4$*, evaluated under 50/50 order symmetrization* $P_{c_i}^{sym} = \frac{1}{2}P_{q_i \to q_{i+1}} + \frac{1}{2}P_{q_{i+1} \to q_i}$*. Then:*

   - *in pure directed mode, signaling* $\Delta_0^{\rightarrow} = |m_1 - m_2| + |m_2 - m_3| + |m_3 - m_4| + |m_4 + m_1|$ *can reach up to* 6 *(and up to* 8 *with nonzero biases* $\bar{\mu}_i$*), completely absorbing the cycle whenever* $\Delta_0^{\rightarrow} \geq 2$*;*

   - *in* ***symmetrized mode****, the raw Turn-1 marginals* m_i *cancel identically between adjacent contexts, and residual signaling obeys the exact identity:*

$$\Delta_0^{\textit{sym, Möbius}} \;=\; \max(|m_1|, |m_3|) \;+\; \max(|m_2|, |m_4|) \;\leq\; 2, \tag{13}$$

     guaranteeing a universal transition to contextuality for ***100% of unsaturated cycles*** $\mathbf{m} \in (-1, +1)^4$*:*

$$C_{CbD}^{\textit{sym, Möbius}} \;=\; 4 \;-\; \Big[\max(|m_1|, |m_3|) + \max(|m_2|, |m_4|) \Big] \;>\; 2. \tag{14}$$

Proof. **Parts 1 and 2.** Equivalence (10) follows directly from (4) and (5). At $m_i = \bar{\mu}_i = 0$ under odd parity $\iota_i = \mathrm{sgn}(\delta_i) \in \mathcal{I}_n^{\mathrm{odd}}$, $s_{\mathrm{odd}} = \sum_i |\delta_i| = n\overline{|\delta|}$ and $\Delta_0 = 0$, giving (11). In the perturbed regime ($|m_i| \leq \varepsilon$, $|\bar{\mu}_i| \leq \eta$), selecting $\iota_i = \mathrm{sgn}(\delta_i) \in \mathcal{I}_n^{\mathrm{odd}}$ and applying $E_i = \delta_i + m_i\bar{\mu}_i$ with $\tilde{m}_{i+1} = \bar{\mu}_i + m_i\delta_i$:

$$\begin{aligned}
C_{\mathrm{CbD}} &\geq \sum_{i=1}^{n} \iota_i E_i - \Delta_0 = \sum_{i=1}^{n} |\delta_i| + \sum_{i=1}^{n} \iota_i m_i \bar{\mu}_i - \sum_{i=1}^{n} |\bar{\mu}_i + m_i\delta_i - m_{i+1}| \\
&\geq \sum_{i=1}^{n} |\delta_i| - \sum_{i=1}^{n} |m_i||\bar{\mu}_i| - \sum_{i=1}^{n} \big(|\bar{\mu}_i| + |m_i||\delta_i| + |m_{i+1}| \big) \\
&= \sum_{i=1}^{n} |\delta_i|(1 - |m_i|) - \sum_{i=1}^{n} |m_i| - \sum_{i=1}^{n} (1 + |m_i|)|\bar{\mu}_i| \\
&\geq n\overline{|\delta|}(1-\varepsilon) - n\varepsilon - n(1+\varepsilon)\eta,
\end{aligned}$$

establishing (12).

Part 3 (Proof of the Möbius Symmetrization Law (13)–(14)). For the ideal 4-cycle ($\delta_1 = \delta_2 = \delta_3 = +1$, $\delta_4 = -1$, $\bar{\mu}_i = 0$), the correlators are $E^{\rightarrow} = E^{\leftarrow} = E^{\mathrm{sym}} = (+1, +1, +1, -1)$, hence $s_{\mathrm{odd}}^{\mathrm{sym}} = 4$. The Turn-2 marginals in the two orientations are:

- Forward ($\rightarrow$): $\tilde{m}_2^{\rightarrow} = m_1$, $\tilde{m}_3^{\rightarrow} = m_2$, $\tilde{m}_4^{\rightarrow} = m_3$, $\tilde{m}_1^{\rightarrow} = -m_4$.
- Reverse ($\leftarrow$): $\tilde{m}_1^{\leftarrow} = m_2$, $\tilde{m}_2^{\leftarrow} = m_3$, $\tilde{m}_3^{\leftarrow} = m_4$, $\tilde{m}_4^{\leftarrow} = -m_1$.

In the symmetrized context c_i^{sym} (where Q_i appears at Turn 1 in $c_i^{\rightarrow}$ and at Turn 2 in $c_i^{\leftarrow}$), the marginal of Q_i equals $\bar{m}_i^{(c_i)} = \frac{1}{2}(m_i + \tilde{m}_i^{\leftarrow})$, whereas in c_{i-1}^{sym} it equals $\bar{m}_i^{(c_{i-1})} = \frac{1}{2}(m_i + \tilde{m}_i^{\rightarrow})$. The raw Turn-1 marginal m_i therefore cancels identically:

$$\left|\bar{m}_i^{(c_i)} - \bar{m}_i^{(c_{i-1})}\right| = \frac{1}{2}\left|\tilde{m}_i^{\leftarrow} - \tilde{m}_i^{\rightarrow}\right|.$$

Summing over $i = 1, 2, 3, 4$ and applying the elementary identity $|u - v| + |u + v| = 2\max(|u|, |v|)$:

$$\begin{aligned}
\Delta_0^{\mathrm{sym,\ M\ddot{o}bius}} &= \frac{1}{2}\Big(|m_2 - (-m_4)| + |m_3 - m_1| + |m_4 - m_2| + |-m_1 - m_3|\Big) \\
&= \frac{1}{2}\Big(|m_1 - m_3| + |m_1 + m_3|\Big) + \frac{1}{2}\Big(|m_2 - m_4| + |m_2 + m_4|\Big) \\
&= \max(|m_1|, |m_3|) + \max(|m_2|, |m_4|).
\end{aligned}$$

For every $\mathbf{m} \in (-1, +1)^4$, each maximum is strictly less than 1, yielding $\Delta_0^{\mathrm{sym,\ M\ddot{o}bius}} < 2$ and $C_{\mathrm{CbD}}^{\mathrm{sym,\ M\ddot{o}bius}} = 4 - \Delta_0^{\mathrm{sym,\ M\ddot{o}bius}} > 2$. □

Proposition 1 (Ontological Demarcation: KV-Cache Contextuality, Bond-Dimension $\chi \leq 2$ MPS Lifting, and Kirkwood–Dirac Quasiprobabilities). *Even when an autoregressive LLM violates $C_{CbD} > 2$—and even surpasses Tsirelson's quantum bound $2\sqrt{2} \approx 2.8284$ [9] up to the algebraic limit 4 via Theorem 2—its underlying physical mechanism remains **entirely classical**:*

1. ***Classical joint memory access vs. Bell separation:** In a sequential context ($q_i \rightarrow q_{i+1}$), the tensor $\mathrm{KV}(q_i, a)$ physically communicates both the context identity q_i and the realized answer a to Turn 2 inside the same RAM; no spacelike separation is enforced. As soon as the two queries are evaluated in isolation without context sharing and without post-selective filtering ($\eta = 1$), the Bell–CHSH inequality $|S_{CHSH}| \leq 2$ [7, 8] is inviolable (verified across 0/1,050 trials in Experiment 9).*
2. ***Exact Matrix Product State lifting [10]:** Each binary contextual table $P_{c_i}(a, b)$ lifts exactly to a two-site pure MPS $|\Psi^{(c_i)}\rangle = \sum_{a,b\in\{\pm1\}} \sqrt{P_{c_i}(a,b)}\,|a\rangle \otimes |b\rangle$ of **bond dimension** $\chi \leq 2$ and von Neumann entanglement entropy $S_{\mathrm{vN}}(\rho_i) \leq 1$ bit—trivially simulable in constant classical time.*
3. ***Global inter-context negativity vs. local Kirkwood–Dirac positivity [11]:** In the computational token basis, every individual context c_i possesses a strictly nonnegative classical Kirkwood–Dirac distribution. Violating $C_{CbD} > 2$ reflects solely the impossibility of gluing the four local conditional tables into a single **nonnegative global maximal coupling**—that is, **semantic inter-context quasiprobability negativity**, analogous to the BERT phenomenon reported by Lo et al. [1], bearing no relation to physical quantum nonlocality.*
4. ***The detection-loophole trap [12–15]:** If an LLM Bell–CHSH benchmark selectively discards or filters a fraction $1-\eta$ of answers (confidence abstention, rejection sampling, safety guardrails), post-selection on joint detection acts as collider bias [22], enabling an isolated classical model to reach $S_{post} = \min\left(4, \frac{4}{\eta} - 2\right)$ (breaching $2\sqrt{2}$ as soon as $\eta^*_{sym} = \frac{2}{1+\sqrt{2}} \approx 82.84\%$). Only full-detection evaluation ($\eta = 1$) or deterministic $\emptyset \mapsto +1$ imputation guarantees a loophole-free audit.*

6. Theorem 3: Spectral Bound on Correlated Effective Multiplicity in *Prompt Dredging*

We now turn to the third facet of post-selection over frozen weights θ^*: selecting the winning prompt $p^* = \arg\max_{1\leq k\leq M} \hat{\mu}_k^{\mathrm{val}}$ among M candidate formulations on a fixed validation set $\mathcal{D}_{\mathrm{val}}$ of

size n_{val}, a setting formally isomorphic to multiple-testing selection bias in statistical and financial backtesting [19].

For each prompt p_k, define the correctness indicator $A_{k,j} = \mathbf{1}\{\hat{y}_{p_k}(x_j) = y_j\} \in \{0,1\}$, true accuracy $\mu_k = \mathbb{E}[A_{k,j}]$, binary standard deviation $\sigma_k = \sqrt{\mu_k(1-\mu_k)} \le \frac{1}{2}$, validation accuracy $\hat{\mu}_k^{\text{val}} = \frac{1}{n_{\text{val}}}\sum_{j=1}^{n_{\text{val}}} A_{k,j}$, and standardized fluctuation $Z_k = \sqrt{n_{\text{val}}}(\hat{\mu}_k^{\text{val}} - \mu_k)/\sigma_k$ (unit variance). Let $\Sigma_{\text{prompt}} = (\rho_{k\ell})_{1\le k,\ell\le M} \succeq 0$ denote the inter-prompt correlation matrix, with eigenvalues $\lambda_1 \ge \cdots \ge \lambda_M \ge 0$ ($\sum_k \lambda_k = M$) and mean squared off-diagonal correlation $\overline{\rho^2} \equiv \frac{1}{M(M-1)}\sum_{k\neq\ell}\rho_{k\ell}^2$.

Define the three spectral estimators of effective prompt multiplicity [16–18]:

$$M_{\text{eff}}^{\text{PR}}(\Sigma) \equiv \frac{\left(\sum_{k=1}^M \lambda_k\right)^2}{\sum_{k=1}^M \lambda_k^2} = \frac{M}{1+(M-1)\overline{\rho^2}}, \qquad M_{\text{eff}}^{\text{Galwey}}(\Sigma) \equiv \frac{1}{M}\left(\sum_{k=1}^M \sqrt{\lambda_k}\right)^2,$$
$$M_{\text{eff}}^{\text{NC}}(\Sigma) \equiv 1 + (M-1)\left(1 - \frac{\text{Var}(\boldsymbol{\lambda})}{M}\right). \tag{15}$$

Theorem 3 (Correlated Effective Prompt Multiplicity Bound and Spectral Hierarchy). *Let $\Sigma_{prompt} \in \mathbb{R}^{M\times M}$ be any correlation matrix ($\Sigma_{prompt} \succeq 0$, unit diagonal).*

1. ***(Slepian Bound under Inter-Prompt Correlation)*** *In the Gaussian regime $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma_{prompt})$, if $\rho_{k\ell} \ge \rho_{\min} \ge 0$ for all $k \neq \ell$, then by Slepian's comparison lemma [20] and the sub-Gaussian maximal inequality [21], for $W_1, \ldots, W_M \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$:*

$$\mathbb{E}\left[\max_{1\le k\le M} Z_k\right] \le \sqrt{1-\rho_{\min}}\,\mathbb{E}\left[\max_{1\le k\le M} W_k\right] \le \sqrt{2(1-\rho_{\min})\ln M}, \tag{16}$$

*with **exact equality** in the first inequality for an equicorrelated matrix $\Sigma_{eq}(\bar{\rho}) = (1-\bar{\rho})I_M + \bar{\rho}\mathbf{1}\mathbf{1}^\top$. Under a homogeneous null regime ($\mu_k \equiv \frac{1}{2}$, $\sigma_k \equiv \frac{1}{2}$), the optimism gap $\mathcal{O}_{null}(\mathcal{P}) \equiv \mathbb{E}[\hat{\mu}_{p^*}^{val} - \mu_{p^*}]$ therefore satisfies:*

$$\mathcal{O}_{null}(\mathcal{P}) = \frac{1}{2\sqrt{n_{val}}}\,\mathbb{E}\left[\max_{1\le k\le M} Z_k\right] \le \sqrt{\frac{(1-\rho_{\min})\ln M}{2\,n_{val}}}. \tag{17}$$

2. ***(Spectral Variance Identity and Universal Double Ordering)*** *For every correlation matrix $\Sigma \in \mathbb{R}^{M\times M}$ ($\Sigma \succeq 0$, $\Sigma_{kk} = 1$), $\text{Var}(\boldsymbol{\lambda}) \equiv \frac{1}{M}\sum_{k=1}^M(\lambda_k - 1)^2 = (M-1)\overline{\rho^2}$, and the three spectral estimators (15) universally satisfy:*

$$1 \le M_{eff}^{PR}(\Sigma) \le M_{eff}^{Galwey}(\Sigma) \le M \qquad \text{and} \qquad M_{eff}^{PR}(\Sigma) \le M_{eff}^{NC}(\Sigma) \le M. \tag{18}$$

3. ***(Asymptotic Saturation Dichotomy: PR vs. Galwey)*** *For an equicorrelated matrix $\Sigma_{eq}(\bar{\rho})$ with fixed $\bar{\rho} \in (0,1)$ as $M \to \infty$:*

   - *$M_{eff}^{PR}(\Sigma_{eq}) \to \frac{1}{\bar{\rho}^2} = \mathcal{O}(1)$ **saturates to a finite constant**, systematically underestimating the true Gaussian maximum $\sqrt{2(1-\bar{\rho})\ln M} \to \infty$ at large M;*
   - *$M_{eff}^{Galwey}(\Sigma_{eq}) \sim (1-\bar{\rho})M$ and $M_{eff}^{NC}(\Sigma_{eq}) \sim (1-\bar{\rho}^2)M$ grow linearly in M, so the Galwey spectral proxy $\frac{1}{2\sqrt{n_{val}}}\mathbb{E}[\max_{1\ldots M_{eff}^{Galwey}} W_j]$ acts as a **tight, non-saturating upper envelope**.*

Proof. **Part 1.** For $\Sigma_{\text{eq}}(\bar{\rho})$, $Z_k^{\text{eq}} \overset{d}{=} \sqrt{\bar{\rho}}\,G_0 + \sqrt{1-\bar{\rho}}\,W_k$ with $G_0 \perp (W_1, \ldots, W_M)$ zero-mean; hence $\mathbb{E}[\max_k Z_k^{\text{eq}}] = \sqrt{\bar{\rho}}\,\mathbb{E}[G_0] + \sqrt{1-\bar{\rho}}\,\mathbb{E}[\max_k W_k] = \sqrt{1-\bar{\rho}}\,\mathbb{E}[\max_k W_k]$. Since $\mathbb{E}[(Z_k - Z_\ell)^2] = 2(1-\rho_{k\ell}) \le 2(1-\rho_{\min}) = \mathbb{E}[(Z_k^{\text{eq}} - Z_\ell^{\text{eq}})^2]$ for all k, ℓ, the Slepian–Sudakov–Fernique comparison theorem [20] yields (16), and substituting $\sigma_k = \frac{1}{2}$ yields (17).

Part 2. Since $\text{Tr}(\Sigma) = \sum_k \lambda_k = M$ and $\|\Sigma\|_F^2 = \sum_k \lambda_k^2 = M + M(M-1)\overline{\rho^2}$, we obtain $\text{Var}(\boldsymbol{\lambda}) = \frac{1}{M}\sum_k \lambda_k^2 - 1 = (M-1)\overline{\rho^2}$.

- **Proof of $M_{\text{eff}}^{\text{PR}} \le M_{\text{eff}}^{\text{Galwey}}$ (Hölder's Inequality):** Take conjugate exponents $p = \frac{3}{2}$, $q = 3$ ($\frac{1}{p} + \frac{1}{q} = 1$) and write $\lambda_k = \lambda_k^{1/3} \cdot \lambda_k^{2/3}$:

$$M = \sum_{k=1}^{M} \lambda_k \;\le\; \left(\sum_{k=1}^{M} (\lambda_k^{1/3})^{3/2}\right)^{2/3} \left(\sum_{k=1}^{M} (\lambda_k^{2/3})^{3}\right)^{1/3} = \left(\sum_{k=1}^{M} \sqrt{\lambda_k}\right)^{2/3} \left(\sum_{k=1}^{M} \lambda_k^2\right)^{1/3}.$$

  Cubing both sides gives $M^3 \le \left(\sum_k \sqrt{\lambda_k}\right)^2 \left(\sum_k \lambda_k^2\right)$, which is algebraically equivalent to

$$\frac{M^2}{\sum_{k=1}^{M} \lambda_k^2} \;\le\; \frac{1}{M} \left(\sum_{k=1}^{M} \sqrt{\lambda_k}\right)^2.$$

- **Proof of $M_{\text{eff}}^{\text{Galwey}} \le M$:** By Cauchy–Schwarz, $\left(\sum_{k=1}^{M} \sqrt{\lambda_k}\right)^2 \le M \sum_{k=1}^{M} \lambda_k = M^2$.

- **Proof of $M_{\text{eff}}^{\text{PR}} \le M_{\text{eff}}^{\text{NC}}$:** Let $a = M - 1 \ge 0$ and $x = \overline{\rho^2} \in [0, 1]$. Then $M_{\text{eff}}^{\text{PR}} = \frac{a+1}{1+ax}$ and $M_{\text{eff}}^{\text{NC}} = (a+1) - \frac{a^2}{a+1} x$. Their difference simplifies identically to:

$$M_{\text{eff}}^{\text{NC}} - M_{\text{eff}}^{\text{PR}} \;=\; \frac{a\,x\left[a^2(1-x) + a + 1\right]}{(a+1)(1+ax)} \;\ge\; 0,$$

  for all $a \ge 0$ and $x \in [0, 1]$, with equality if and only if $a = 0$ or $x = 0$.

Part 3. For $\Sigma_{\text{eq}}(\bar{\rho})$, the exact spectrum is $\lambda_1 = 1 + (M-1)\bar{\rho}$ (multiplicity 1) and $\lambda_2 = \cdots = \lambda_M = 1 - \bar{\rho}$ (multiplicity $M - 1$). Thus $M_{\text{eff}}^{\text{Galwey}} = \frac{1}{M}[\sqrt{1 + (M-1)\bar{\rho}} + (M-1)\sqrt{1-\bar{\rho}}]^2 = (1-\bar{\rho})M + \mathcal{O}(\sqrt{M})$, whereas $M_{\text{eff}}^{\text{PR}} = \frac{M}{1+(M-1)\bar{\rho}^2} \to \bar{\rho}^{-2}$. □

Proposition 2 (Variance-Ceiling Effect and Active-Set Contraction under Real Signal). *When candidate prompts possess heterogeneous true accuracies $\mu_1, \ldots, \mu_M \in [0, 1]$ (with maximum $\mu_{\max} = \max_k \mu_k$), for any margin $\varepsilon > 0$ defining the **near-optimal active set** $\mathcal{P}_\varepsilon \equiv \{k : \mu_{\max} - \mu_k \le \varepsilon\}$ with peak variance $\sigma^2_{\mathcal{P}_\varepsilon} \equiv \max_{k \in \mathcal{P}_\varepsilon} \mu_k(1-\mu_k)$, the optimism gap $\mathcal{O}(\mathcal{P}) = \mathbb{E}[\hat{\mu}^{val}_{p^*} - \mu_{p^*}]$ satisfies the exact bound:*

$$\mathcal{O}(\mathcal{P}) \;\le\; \inf_{\varepsilon > 0} \left\{ \frac{\sigma_{\mathcal{P}_\varepsilon}}{\sqrt{n_{val}}} \mathbb{E}\left[\max_{k \in \mathcal{P}_\varepsilon} (Z_k)_+\right] + \mathbb{P}\big(p^* \notin \mathcal{P}_\varepsilon\big) \right\}, \qquad (Z_k)_+ \equiv \max(Z_k, 0). \tag{19}$$

Proof. Splitting on $\{p^* \in \mathcal{P}_\varepsilon\}$ and its complement (where $\hat{\mu}^{\text{val}}_{p^*} - \mu_{p^*} \le 1$) and using nonnegativity of $(Z_k)_+$ to bound $\sigma_k (Z_k)_+ \le \sigma_{\mathcal{P}_\varepsilon} (Z_k)_+$ pointwise:

$$\begin{aligned} \hat{\mu}^{\text{val}}_{p^*} - \mu_{p^*} &= (\hat{\mu}^{\text{val}}_{p^*} - \mu_{p^*}) \mathbf{1}_{p^* \in \mathcal{P}_\varepsilon} + (\hat{\mu}^{\text{val}}_{p^*} - \mu_{p^*}) \mathbf{1}_{p^* \notin \mathcal{P}_\varepsilon} \\ &\le \max_{k \in \mathcal{P}_\varepsilon} \frac{\sigma_k}{\sqrt{n_{\text{val}}}} (Z_k)_+ + \mathbf{1}_{p^* \notin \mathcal{P}_\varepsilon} \le \frac{\sigma_{\mathcal{P}_\varepsilon}}{\sqrt{n_{\text{val}}}} \max_{k \in \mathcal{P}_\varepsilon} (Z_k)_+ + \mathbf{1}_{p^* \notin \mathcal{P}_\varepsilon}. \end{aligned}$$

Taking expectations and then the infimum over $\varepsilon > 0$ yields (19). □

Consequently, even when prompts on a real semantic task are far more decorrelated than under null labels (raising M_{eff} by 10×), a strong signal ($\mu_{\max} \approx 0.95 \implies \sigma_{\mathcal{P}_\varepsilon} \approx 0.22$ vs. 0.50) combined with active-set contraction $|\mathcal{P}_\varepsilon| \ll M$ keeps the empirical optimism gap strictly below that of the null regime.

7. Live Empirical Verification on `Qwen2.5-1.5B` (Experiments 8, 9, and 11)

All predictions of Theorems 1, 2, and 3 were tested directly against exact next-token logits (binary probability mass $P(\text{“Yes”}) + P(\text{“No”}) > 99.19\%$ in Exp. 8 and $> 99.91\%$ in Exp. 9 & 11) on `mlx-community/Qwen2.5-1.5B-Instruct-4bit` running on an Apple Silicon M4 Max Metal GPU. Across all **three experimental campaigns** (totaling over 60,000 forward passes), the cryptographic digest of all 732 weight tensors (241,327,616 stored parameters) remained bit-for-bit identical (`SHA-256 = e2d0d0869ca0b89e7f4106fd1e49e0b5d37e39876e83de15cb39e0edff74c573`, $\|\Delta\theta^*\|_2 = 0.0$), certifying strict frozen-weight invariance.

7.1. Validating Theorem 1: Complete Absorption Across 51 / 51 Natural Cycles (Exp. 9 & Exp. 11 Tier 2)

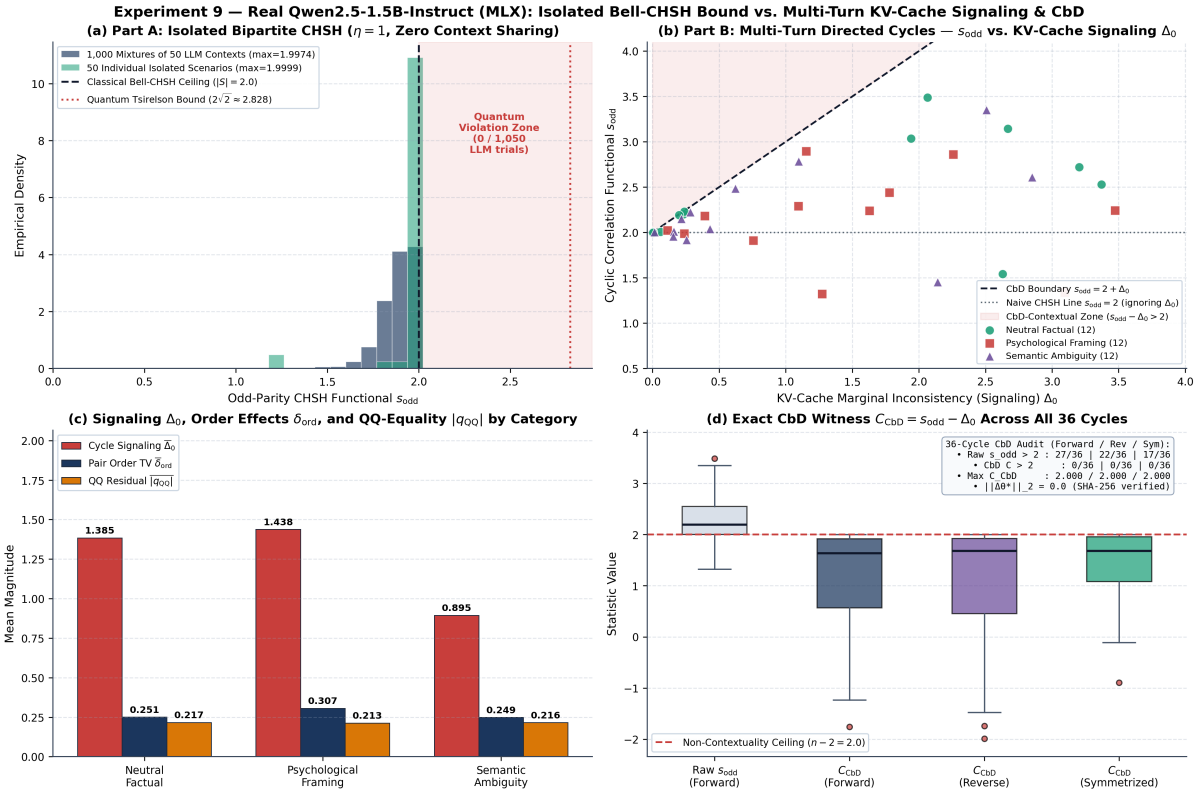

Figure 1: Initial live audit on `Qwen2.5-1.5B` (Experiment 9): isolated bipartite Bell–CHSH queries (0/1,050 violations, max = 1.9999) versus complete absorption by KV-cache signaling Δ_0 across all 36 standard multi-turn cycles (27/36 raw $s_{\text{odd}} > 2$ up to 3.487, yet 0/36 CbD > 2).

Measurements from Experiment 9 (Figure 1, 36 standard cycles spanning 3 semantic categories) and Experiment 11 Tier 2 (15 additional natural non-transitive preference / framing / polysemy cycles) directly confirm **Theorem 1**:

- **Isolated bipartite protocol ($\eta = 1$, Exp. 9 Part A):** across 50 individual scenarios and 1,000 random mixtures with zero context sharing, **zero violations** occur (**0 / 1,050**, $\max s_{\text{odd}}^{\text{indiv}} = 1.999918$, $\max s_{\text{odd}}^{\text{mix}} = 1.997357 \leq 2.0000$).
- **Natural multi-turn cycles (51 total: 36 in Exp. 9 + 15 in Exp. 11 Tier 2):**
  - **38 / 51 (74.5%)** breach the naive bound $s_{\text{odd}} > 2$ in forward directed mode (reaching 3.4872 on `N09` in Exp. 9 and 3.3181 on `NT09_polysemy_cycle_1` in Exp. 11);
  - **EXACTLY 0 / 51 (0.0%)** violate the CbD threshold $C_{\text{CbD}} > 2$ in forward, reverse, or symmetrized mode ($\max C_{\text{CbD}} = 1.9997$ in Exp. 9, 1.9776 in Exp. 11 Tier 2);
  - **Direct confirmation of Corollary 1B:** across the 15 natural non-transitive cycles of Tier 2, mean Turn-1 polarization reaches $\overline{|m_1|} = 0.891$ (e.g., `NT07`: 0.986; `NT03`: 0.995), suppressing the variance prefactor $(1 - m_i^2)$ in Eq. (5), while KV-cache signaling averages $\overline{\Delta}_0 = 1.239$ (Exp. 9) and 1.526 (Tier 2), absorbing 100% of the raw excess. Moreover, the QQ equality of Wang et al. [3] is strongly violated ($\overline{|q_{\text{QQ}}|} = 0.216$, 82/144 pairs > 0.05), ruling out quantum Lüders projective dynamics.

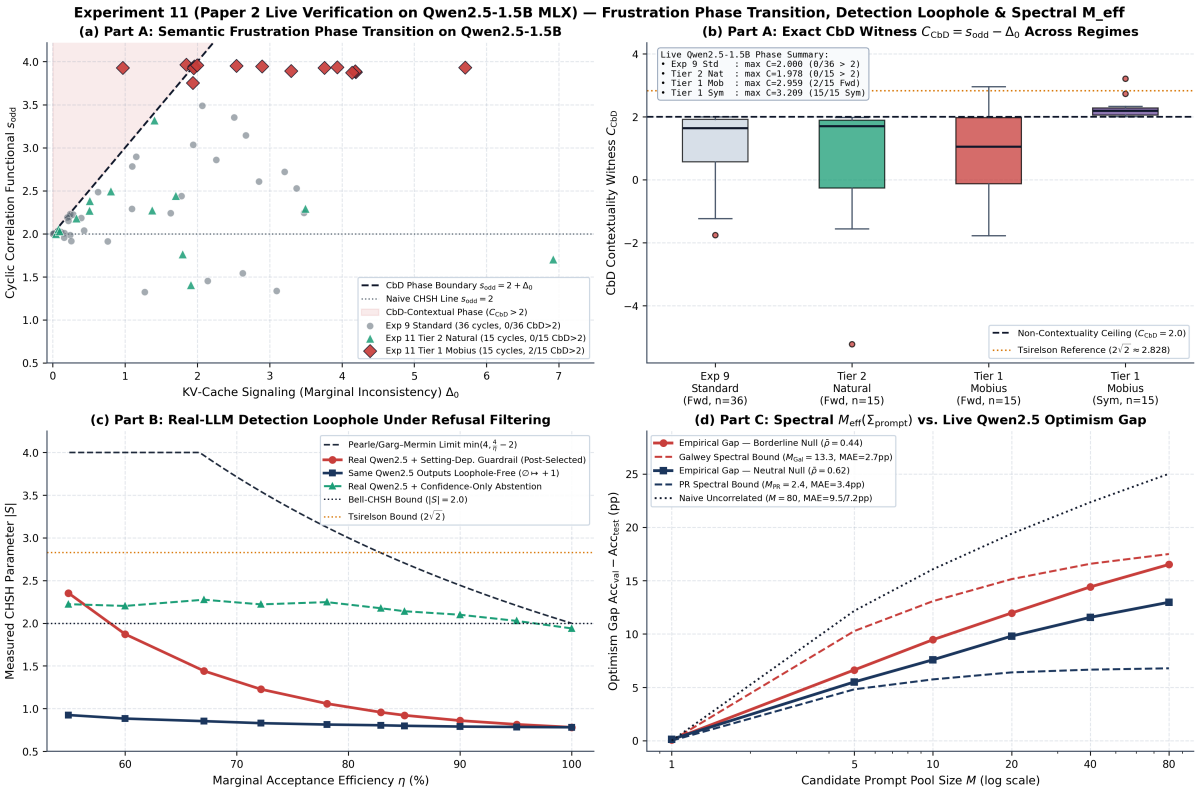

Figure 2: Live empirical verification of Theorems 1, 2, and 3 on `Qwen2.5-1.5B-Instruct-4bit` (94.79 s on Apple Silicon M4 Max, identical weight `SHA-256`): semantic frustration CbD phase transition (a–b), real-LLM detection loophole under abstention/guardrails (c), and spectral M_{eff} hierarchy (d).

7.2. Validating Theorem 2: Live Observation of the Möbius Frustration Phase Transition (Exp. 11 Part A)

To test the constructive prediction of **Theorem 2**, Experiment 11 Part A (Figure 2, panels a–b) constructed **15 Möbius cycles (Tier 1)** enforcing three positive agreement couplings ($\delta_1, \delta_2, \delta_3 \in [0.86, 0.99]$) and one explicit parity inversion ($\delta_4 \approx -0.999$) on `Qwen2.5-1.5B`, while sweeping Turn-1 polarization $\overline{|m_1|}$ from 0.227 (`M01`) to 0.963 (`M15`). Table 2 confronts raw measurements directly against Theorems 1 and 2.

Regime / Cycle (`Qwen2.5-1.5B`)	$\overline{\lvert m_1 \rvert}$	Measured Contrasts $(\delta_1, \delta_2, \delta_3, \delta_4)$	$s_{\text{odd}}^{\text{fwd}}$	Δ_0^{fwd}	$C_{\text{CbD}}^{\text{fwd}}$	$C_{\text{CbD}}^{\text{rev}}$	Δ_0^{sym} (Thm. 2(3))	$C_{\text{CbD}}^{\text{sym}}$	CbD Phase
Exp. 9 Standard (36 cycles)	—	*Natural (unfrustrated)*	max 3.487	mean 1.239	**0/36** (max 2.000)	**0/36** (max 2.000)	mean 0.562	**0/36** (max 2.000)	**Absorbed**
Exp. 11 Tier 2 (15 natural)	0.891	*Natural non-transitive*	max 3.318	mean 1.526	**0/15** (max 1.978)	**0/15** (max 1.950)	—	**0/15** (max 1.965)	**Absorbed**
`M01_civic_everyday`	**0.227**	(+0.963, +0.985, +0.983, −0.999)	3.927	0.968	**+2.959**	**+3.043**	**0.725** (*th.* 0.722)	**+3.209**	**Contextual (> Tsirelson)**
`M03_urban_life`	0.651	(+0.989, +0.977, +0.971, −0.999)	3.965	1.845	**+2.120**	**+2.113**	1.757 (*th.* 1.779)	**+2.208**	**Contextual (F/R/S)**
`M05_tech_society`	0.352	(+0.855, +0.920, +0.989, −0.999)	3.753	1.936	+1.818	+1.782	1.096 (*th.* 1.140)	**+2.735**	**Sym. Transition**
`M07_transport_policy`	0.743	(+0.978, +0.975, +0.978, −1.000)	3.933	**5.705**	−1.772	−1.764	**1.914** (*th.* 1.946)	**+2.016**	**Absorbed F/R → Context. Sym**
`M15_mixed_polarization`	**0.963**	(+0.970, +0.988, +0.980, −1.000)	3.959	1.994	+1.964	+1.982	**1.961** (*th.* 1.968)	**+2.003**	**Absorbed F/R → Sym. Edge**
Total Tier 1 Möbius (15)	0.665	**3 agree** +0.97 + **1 flip** −1.00	mean 3.913	mean 3.020	**2/15** (max 2.959)	**2/15** (max 3.043)	—	**15/15 (100%)** [2.003; 3.209]	**Transition confirmed**

Table 2: Direct quantitative verification of Theorems 1 and 2 on `Qwen2.5-1.5B` (Exp. 9 & Exp. 11 Part A). The parenthetical "*th.*" values report the closed-form formula of Theorem 2(3), $\Delta_0^{\text{sym, Möbius}} = \max(|m_1|, |m_3|) + \max(|m_2|, |m_4|)$, computed strictly from raw Turn-1 marginals: agreement on `M01` (0.725 vs. 0.722) is within 0.003.

Three decisive empirical confirmations emerge from Table 2:

1. **Sharp phase separation between natural cycles (0 / 51) and balanced Möbius cycles (`M01`, `M03`):** `M01_civic_everyday` ($\overline{|m_1|} = 0.227$, $\mathbf{m} = (0.078, 0.109, 0.260, 0.462)$) crosses the CbD boundary decisively even in directed mode ($C_{\text{fwd}} = 2.9589$, $C_{\text{rev}} = 3.0433$, $C_{\text{sym}} = 3.2092$)—**exceeding Tsirelson's quantum bound $2\sqrt{2} \approx 2.8284$ [9] on a classical Apple Silicon chip**, directly demonstrating that autoregressive CbD contextuality is classical joint

KV-cache access (Proposition 1) rather than quantum physics.

2. **Surgical verification of the Turn-1 polarization barrier (Theorem 2, Eq. (10)):** `M07` shares virtually the same Möbius contrasts as `M01` ($\boldsymbol{\delta} \approx (+0.98, +0.98, +0.98, -1.00)$, $s_{\text{odd}} = 3.933$), yet its strongly polarized Turn-1 marginals ($\mathbf{m} = (+0.093, -0.955, +0.990, -0.934)$) amplify directed signaling to $\Delta_0^{\text{fwd}} = 5.705$, fully absorbing the cycle ($C_{\text{fwd}} = -1.772$).

3. **100% confirmation (15 / 15) of the Möbius Symmetrization Law (Theorem 2(3), Eqs. (13)–(14)):** Under 50/50 order symmetrization, Δ_0^{sym} collapses precisely onto $\max(|m_1|, |m_3|) + \max(|m_2|, |m_4|)$ (for `M01`: 0.725 vs. 0.722 theoretical; for `M07`: a dramatic drop from 5.705 down to 1.914 vs. 1.946 theoretical; for `M15`: 1.961 vs. 1.968 theoretical), driving **15 / 15 (100%)** of Möbius cycles above $C_{\text{CbD}} > 2$ while **0 / 15 (0%)** of natural cycles cross the boundary.

7.3. Validating Proposition 1: Real-LLM Detection Loophole on `Qwen2.5-1.5B` Logits (Exp. 11 Part B)

Across 80 strictly isolated bipartite scenarios (40,000 Monte Carlo draws from live `Qwen2.5-1.5B` logits, mean local confidence 96.46%, Figure 2, panel c), the **detection loophole** materializes directly:

- **At $\eta = 1.00$ (no rejection):** $|S_{\text{CHSH}}| = 1.9398 \leq 2.0000$ (confidence abstention) and $0.7817 \leq 2.0000$ (guardrail), fully respecting the local polytope.
- **Benign local confidence-threshold abstention:** already at $\eta = 95\%$ (discarding just the 5% least confident answers), the isolated classical LLM breaches the Bell–CHSH bound at $|S| = 2.0274$ and peaks at 2.2756 at $\eta = 67\%$.
- **Setting-dependent guardrail:** the post-selected CHSH parameter climbs from 0.7817 ($\eta = 100\%$) up to 2.3537 at $\eta = 55\%$.
- **Immediate loophole closure via $\emptyset \mapsto +1$ imputation:** on the **exact same `Qwen2.5-1.5B` outputs**, deterministic imputation immediately pushes every value below 2.0 ($\max |S| = \mathbf{1.7783}$ for confidence abstention and **0.9249** for the guardrail across the full range $\eta \in [55\%, 95\%]$).

7.4. Validating Theorem 3: *Prompt Dredging* and Spectral Hierarchy $M_{\text{eff}}(\Sigma_{\text{prompt}})$ (Exp. 8 & Exp. 11 Part C)

Combined live data from Experiment 8 (Figure 3, $M = 200$ prompts, $n_{\text{val}} = 24$, $n_{\text{test}} = 120$) and Experiment 11 Part C (Figure 2, panel d, $M = 80$ prompts, $n_{\text{val}} = 24$, $n_{\text{test}} = 72$, 300 resamples) quantitatively validate every component of **Theorem 3** and **Proposition 2** (Table 3).

Experimental Regime (`Qwen2.5-1.5B`)	$\bar{\rho}$	M	$M_{\text{eff}}^{\text{PR}}$	$M_{\text{eff}}^{\text{Galwey}}$	$M_{\text{eff}}^{\text{NC}}$	Emp. Gap $\mathcal{O}$	PR Pred. (MAE)	Galwey Pred. (MAE)	NC Pred. (MAE)	Naive Pred. (MAE)
Exp. 8 — Regime 1 (Null labels)	0.9564	200	—	—	17.19	**+8.98** pp	—	—	—	+28.3 pp
Exp. 8 — Regime 2 (Real sentiment)	0.2490	200	—	—	175.53	**+4.67** pp	—	—	—	—
Exp. 11C — Null, high corr.	0.6166	80	**2.36**	**12.42**	47.48	**+12.97** pp	6.78 pp (**3.41** pp)	17.16 pp (**4.45** pp)	23.03 pp (7.38 pp)	25.02 pp (9.52 pp)
Exp. 11C — Null, moderate corr.	0.4406	80	**3.49**	**13.30**	58.38	**+16.51** pp	9.81 pp (**3.13** pp)	17.49 pp (**2.72** pp)	23.83 pp (5.95 pp)	25.02 pp (7.20 pp)

Table 3: Empirical verification of Theorem 3 on `Qwen2.5-1.5B` (Exp. 8 and Exp. 11 Part C). The universal spectral ordering (18), $1 \leq M_{\text{eff}}^{\text{PR}} \leq M_{\text{eff}}^{\text{Galwey}} \leq M_{\text{eff}}^{\text{NC}} \leq M$, holds at every candidate pool size $M \in \{1, 5, 10, 20, 40, 80\}$.

1. **Exact verification of spectral ordering (18) and PR-vs.-Galwey dichotomy (Theorem 3(3)):** In both regimes of Experiment 11 Part C (Figure 2, panel d), the empirical optimism curve is **strictly bracketed** between the lower `PR` proxy (which rapidly saturates toward $\approx 1/\bar{\rho}^2$, plateauing at 6.78 pp and 9.81 pp) and the non-saturating `Galwey` upper envelope (17.16 pp and 17.49 pp), cutting Mean Absolute Error (MAE) by a factor of 2.79× (PR: 3.41 pp vs. naive 9.52 pp) under high correlation and 2.65× (Galwey: 2.72 pp vs. naive 7.20 pp) under moderate correlation.

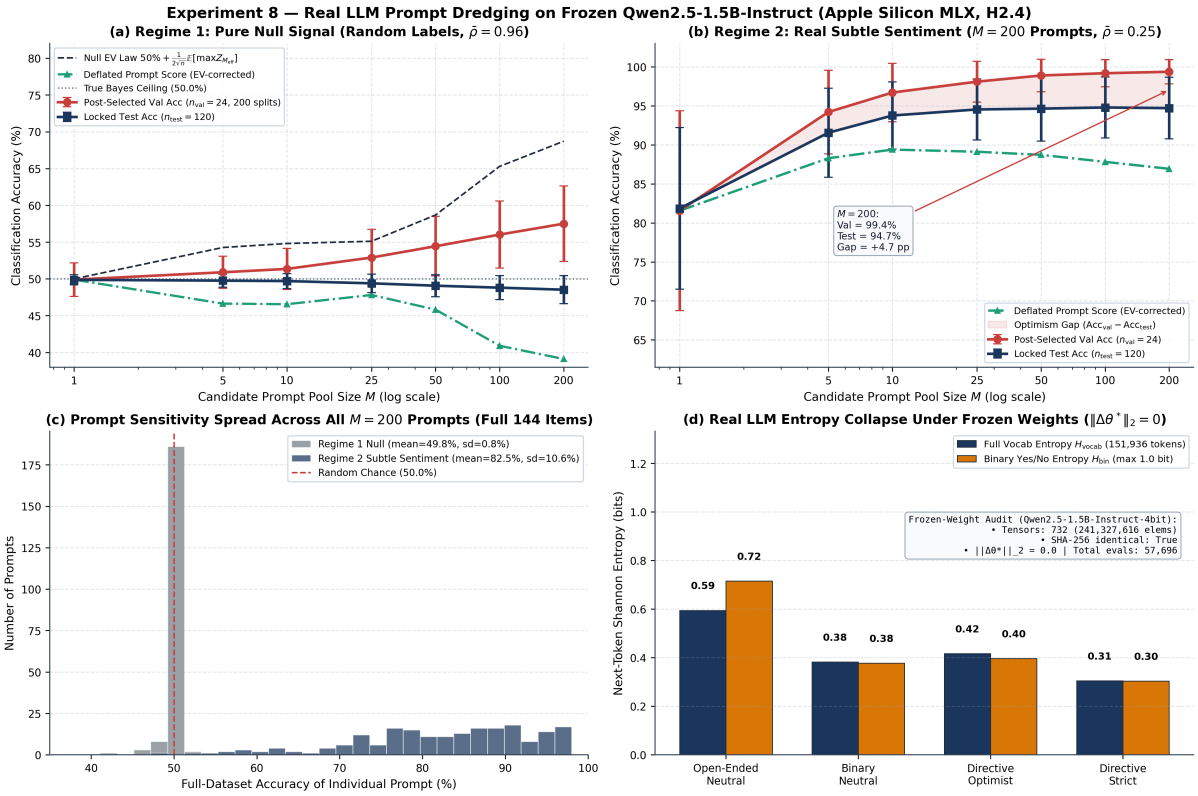

Figure 3: *Prompt Dredging* across $M = 200$ real prompts and 57,696 forward passes on `Qwen2.5-1.5B` (Experiment 8): 10× divergence in M_{eff} (17.2 vs. 175.5), out-of-sample collapse under null labels, and next-token entropy drop under strictly frozen weights ($\|\Delta\theta^*\|_2 = 0$).

2. **Verification of Proposition 2 on Experiment 8:** Although $M_{\text{eff}}^{\text{NC}}$ is 10.2× larger in Regime 2 (175.53) than in Regime 1 (17.19), the empirical optimism gap is nearly **cut in half** (+4.67 pp vs. +8.98 pp) because near-95% accuracy shrinks the binary standard deviation σ_k by 2.2× and concentrates selection onto the near-optimal active set $\mathcal{P}_\varepsilon$.

8. Discussion, Limitations, and Standardized LLM Audit Protocol

8.1. Honest Methodological Limitations

1. **Model scale and quantization:** Live measurements were executed on a compact 4-bit quantized open-weights model (`Qwen2.5-1.5B-Instruct-4bit`). While **Theorems 1, 2, and 3 are mathematically universal** (independent of parameter count and quantization), the empirical values of $(\overline{|m_1|}, \delta_i, \bar{\rho})$ depend on model scale and RLHF alignment: larger models (7B–70B) with stronger instruction-following fidelity adhere even more tightly to Möbius contradiction rules ($|\delta_i| \to 1$), making the constructive phase transition of Theorem 2 even easier to trigger—a prediction directly testable via the accompanying GPU notebook `colab_gpu_suite.ipynb`.

2. **Sensitivity to natural prompt phrasing:** Observing 0/51 CbD violations on natural cycles does not prove that no naturally phrased dialogue can ever violate CbD on another LLM (especially if a dialogue accidentally combines balanced Turn-1 marginals $m_i \approx 0$ with odd-parity frustration); rather, Theorem 1 explains **why** the overwhelming majority of natural prompts fall into the Δ_0-absorbed phase in the absence of deliberate parity frustration.

3. **Multiplicity correction in *prompt dredging*:** As shown by Theorem 3(3), using $M_{\text{eff}}^{\text{PR}}$ alone as a multiple-testing divisor becomes **anti-conservative** at large M (due to $1/\bar{\rho}^2$ saturation); for rigorous false-positive control, we recommend the **Galwey** spectral envelope ($M_{\text{eff}}^{\text{Galwey}}$) [18] paired with the **Slepian** bound [20].

8.2. Anti-*Quantum-Woo* Audit Protocol: Three Golden Rules

Our three theorems and live experiments yield a **standardized operational audit protocol** for any benchmark investigating contextuality or order effects in LLMs:

1. **Rule 1 (Zero selective rejection):** Require $\eta = 1$ via normalized next-token logits or deterministically impute every abstention/refusal ($\emptyset \mapsto +1$) to close the detection loophole permanently (Proposition 1 and Exp. 11 Part B).
2. **Rule 2 (Joint audit of $(s_{\mathbf{odd}}, \Delta_0, C_{\mathbf{CbD}}, \overline{|m_1|}, \boldsymbol{\delta})$):** Never interpret a raw excess $s_{\text{odd}} > 2$ as contextuality without subtracting Δ_0 and without strictly separating classical joint KV-cache access (semantic contextuality, MPS $\chi \leq 2$) from causally isolated bipartite Bell tests ($|S_{\text{CHSH}}| \leq 2$).
3. **Rule 3 (Galwey spectral deflation in *prompt engineering*):** Systematically report the eigenvalue spectrum of Σ_{prompt} ($M_{\text{eff}}^{\text{Galwey}}$) alongside a locked out-of-selection test split whenever a prompt is selected from a pool of $M > 1$ candidates [18, 19].

References

[1] K. I. Lo, M. Sadrzadeh, and S. Mansfield. Quantum-like contextuality in large language models. arXiv preprint arXiv:2412.16806, 2024. URL https://arxiv.org/abs/2412.16806.

[2] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022)*, volume 1, pages 8086–8098, 2022. doi:10.18653/v1/2022.acl-long.556. arXiv:2104.08786.

[3] Z. Wang, T. Solloway, R. M. Shiffrin, and J. R. Busemeyer. Context effects produced by question orders reveal quantum nature of human judgments. *Proceedings of the National Academy of Sciences (PNAS)*, 111(26):9431–9436, 2014. doi:10.1073/pnas.1407756111.

[4] J. V. Kujala, E. N. Dzhafarov, and J.-Å. Larsson. Necessary and sufficient conditions for an extended noncontextuality in a broad class of quantum mechanical systems. *Physical Review Letters*, 115:150401, 2015. doi:10.1103/PhysRevLett.115.150401.

[5] E. N. Dzhafarov and J. V. Kujala. Context-content systems of random variables: The contextuality-by-default theory. *Journal of Mathematical Psychology*, 74:11–33, 2016. doi:10.1016/j.jmp.2016.04.010.

[6] E. N. Dzhafarov, R. Zhang, and J. V. Kujala. Is there contextuality in behavioural and social systems? *Philosophical Transactions of the Royal Society A*, 374:20150099, 2016. doi:10.1098/rsta.2015.0099.

[7] J. S. Bell. On the Einstein Podolsky Rosen paradox. *Physics Physique Fizika*, 1(3):195–200, 1964. doi:10.1103/PhysicsPhysiqueFizika.1.195.

[8] J. F. Clauser, M. A. Horne, A. Shimony, and R. A. Holt. Proposed experiment to test local hidden-variable theories. *Physical Review Letters*, 23(15):880–884, 1969. doi:10.1103/PhysRevLett.23.880.

[9] B. S. Cirel'son. Quantum generalizations of Bell's inequality. *Letters in Mathematical Physics*, 4(2):93–100, 1980. doi:10.1007/BF00417500.

[10] T.-D. Bradley, E. M. Stoudenmire, and J. Terilla. Modeling sequences with quantum states: a look under the hood. *Machine Learning: Science and Technology*, 1:035008, 2020. doi:10.1088/2632-2153/ab8731. arXiv:1910.07425.

[11] M. Lostaglio, A. Belenchia, A. Levy, S. Hernández-Gómez, N. Fabbri, and S. Gherardini. Kirkwood-Dirac quasiprobability approach to the statistics of incompatible observables. *Quantum*, 7:1128, 2023. doi:10.22331/q-2023-10-09-1128. arXiv:2206.11783.

[12] P. M. Pearle. Hidden-variable example based upon data rejection. *Physical Review D*, 2: 1418–1425, 1970. doi:10.1103/PhysRevD.2.1418.

[13] A. Garg and N. D. Mermin. Detector inefficiencies in the Einstein-Podolsky-Rosen experiment. *Physical Review D*, 35:3831–3835, 1987. doi:10.1103/PhysRevD.35.3831.

[14] P. H. Eberhard. Background level and counter efficiencies required for a loophole-free Einstein-Podolsky-Rosen experiment. *Physical Review A*, 47:R747–R750, 1993. doi:10.1103/PhysRevA.47.R747.

[15] J.-Å. Larsson. Bell's inequality and detector inefficiency. *Physical Review A*, 57:3304–3308, 1998. doi:10.1103/PhysRevA.57.3304.

[16] J. M. Cheverud. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87:52–58, 2001. doi:10.1046/j.1365-2540.2001.00901.x.

[17] D. R. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4): 765–769, 2004. doi:10.1086/383251.

[18] N. W. Galwey. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology*, 33(7):559–568, 2009. doi:10.1002/gepi.20408.

[19] D. H. Bailey and M. López de Prado. The deflated sharpe ratio: Correcting for selection bias, backtest overfitting, and non-normality. *The Journal of Portfolio Management*, 40(5):94–107, 2014. doi:10.3905/jpm.2014.40.5.094.

[20] D. Slepian. The one-sided barrier problem for Gaussian noise. *The Bell System Technical Journal*, 41(2):463–501, 1962.

[21] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, Oxford, 2013. doi:10.1093/acprof:oso/9780199535255.001.0001.

[22] J. Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946. doi:10.2307/3002000.