COLLECTIVE ALIGNMENT IN LLM MULTI-AGENT SYSTEMS: DISENTANGLING BIAS FROM COOPERATION VIA STATISTICAL PHYSICS

Anonymous authors
Paper under double-blind review

ABSTRACT

We investigate the emergent collective dynamics of LLM-based multi-agent systems on a 2D square lattice and present a model-agnostic statistical-physics method to disentangle social conformity from intrinsic bias, compute critical exponents, and probe the collective behavior and possible phase transitions of multi-agent systems. In our framework, each node of an $L \times L$ lattice hosts an identical LLM agent holding a binary state ($+1/-1$, mapped to `yes`/`no`) and updating it by querying the model conditioned on the four nearest-neighbor states. The sampler temperature T serves as the sole control parameter. Across three open-weight models (`llama3.1:8b`, `phi4-mini:3.8b`, `mistral:7b`), we measure magnetization and susceptibility under a global-flip protocol designed to probe $\mathbb{Z}_2$ symmetry. All models display temperature-driven order-disorder crossovers and susceptibility peaks; finite-size scaling on even-L lattices yields effective exponents γ/ν that are model-dependent and, for `llama3.1:8b` and `mistral:7b`, manifestly incompatible with the 2D Ising universality class ($\gamma/\nu = 7/4$). Our method enables the extraction of effective β-weighted couplings $\tilde{J}(T)$ and fields $\tilde{h}(T)$, which serve as a measure of social conformity and intrinsic bias. In the models we analyzed, we found that collective alignment is dominated by an intrinsic bias ($\tilde{h} \gg \tilde{J}$) rather than by cooperative neighbor coupling, producing field-driven crossovers instead of genuine phase transitions. These effective parameters vary qualitatively across models, providing compact collective-behavior fingerprints for LLM agents and a quantitative diagnostic for the reliability of multi-agent consensus and collective alignment.

1 INTRODUCTION

Multi-agent systems (MAS) of large language models are rapidly becoming a foundational architecture for collaborative and collective AI Guo et al. (2024); Hong et al. (2024), with applications spanning automated debate Du et al. (2023); Smit et al. (2024), scientific reasoning Wang et al. (2022), content moderation Ruan et al. (2025), collective decision-making Chuang et al. (2024); Piatti et al. (2024), agents economy Tomasev et al. (2025); Wang et al. (2026), and decentralized AI Sun et al. (2023); Ding et al. (2025); Riehl et al. (2026). As these systems scale from research prototypes to production deployments, a fundamental question becomes urgent: when a group of LLM agents reaches apparent consensus, is this alignment the result of genuine cooperation, or simply the amplification of shared biases encoded in the underlying model? The answer has direct consequences for MAS alignment and for the reliability of MAS-based pipelines, where false consensus can propagate errors and suppress diversity of reasoning.

Statistical physics offers a natural and powerful language for these phenomena. Order parameters quantify the degree of collective alignment, control parameters tune the balance between order and disorder, and universality classes classify macroscopic behavior independently of microscopic details. The 2D Ising model is the simplest nontrivial system exhibiting a continuous phase transition with well-characterized critical exponents, making it an ideal theoretical reference point. By placing LLM agents on a square lattice with binary states and nearest-neighbor interactions, we can ask directly whether LLM collectives fall into the Ising universality class or depart from it, and, crucially, *why*.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The central contribution of this work is a model-agnostic statistical-physics method for studying multi-agent systems. The method takes as input only the observable update logs of any LLM that can be queried with a structured prompt, requiring no access to the model's internal weights, logits, or architecture. From these logs alone, the method enables: (i) the disentangling of intrinsic single-agent bias from cooperative inter-agent influence through an effective coupling $\tilde{J}$ and field $\tilde{h}$ extracted via logistic regression; (ii) the computation of critical exponents through standard finite-size scaling of canonical observables such as the magnetization and the susceptibility; and (iii) the broader exploration of statistical and collective behavior, including the diagnosis of phase transitions and crossover phenomena, in arbitrary multi-agent architectures.

In this work, we apply this method to a 2D square-lattice setup across three open-weight models, providing the first systematic finite-size study of LLM agents on a lattice. We measure canonical observables (magnetization $m(T)$ and susceptibility $\chi(T)$), test finite-size scaling against 2D Ising predictions, and extract $(\tilde{J}, \tilde{h})$ trajectories that decompose collective behavior into "social conformity" (measured by $\tilde{J}$) and "intrinsic bias" (measured by $\tilde{h}$). In the language of recent alignment work, $\tilde{h}$ might be the lattice-scale analog of sycophancy and label bias, while $\tilde{J}$ measures genuine inter-agent influence. We show that the dominant mechanism behind low-temperature alignment is the effective field, not cooperative coupling, and that the resulting crossover phenomenology is model-dependent and generically non-Ising.

Beyond characterizing collective dynamics, the $(\tilde{J}, \tilde{h})$ decomposition serves as a probe of the LLM itself. Each model traces a distinct trajectory in $(\tilde{J}, \tilde{h})$ space as temperature varies, revealing how strongly it follows its neighbors versus its own prior. These trajectories act as "collective-behavior fingerprints" that expose properties of the model (such as the strength of its label bias and its responsiveness to context) that would be difficult to measure through single-agent benchmarks alone.

These results bear directly on multi-agent alignment. Many MAS architectures (debate, self-consistency, committee voting) implicitly assume that agreement among multiple agents provides stronger evidence for correctness than a single query. However, when all agents are copies of the same model, they share the same intrinsic biases, encoded in their weights Santagata & De Nobili (2024); Zheng et al. (2024) and shaped by alignment training such as RLHF Christiano et al. (2017); Ouyang et al. (2022). A well-documented manifestation of these training-induced biases is sycophancy Sharma et al. (2023), the tendency of preference-tuned models to align responses with user-suggested or context-suggested positions; in our framework, this and related label preferences appear as a nonzero effective field. If these shared biases dominate over genuine inter-agent influence (the $\tilde{h} \gg \tilde{J}$ regime identified in this work), then N agents agreeing carries no more evidential weight than a single agent responding N times: the consensus is correlated through the shared bias, not produced by independent deliberation. The $(\tilde{J}, \tilde{h})$ decomposition provides a quantitative pre-deployment test of whether a given model, prompt, and temperature combination falls in this bias-dominated regime or in the genuinely cooperative regime where collective agreement is informative. Furthermore, the susceptibility peak identifies the temperature at which the collective is most sensitive to perturbation, a quantity that is directly useful for designing human-in-the-loop MAS protocols: at this temperature, a human overseer can most effectively steer the group's opinion. Finally, evidence of non-equilibrium dynamics in some models (Sec. 4.1 and Appendix C) suggests that standard game-theoretic and statistical-equilibrium frameworks, on which much of alignment theory relies, may not fully capture the behavior of LLM collectives, motivating the development of new theoretical tools at the intersection of non-equilibrium physics and alignment.

Our work connects to several active research directions: LLM-based debate and self-consistency protocols Wang et al. (2022); Du et al. (2023); Smit et al. (2024); LLM-as-judge evaluation and its known biases (position, verbosity, self-enhancement) Zheng et al. (2023); Li et al. (2024); alignment via RLHF and the systematic preferences it induces, including sycophancy Christiano et al. (2017); Ouyang et al. (2022); Sharma et al. (2023); opinion-dynamics models on networks (voter model, majority rule, bounded confidence) Castellano et al. (2009); Mullick & Sen (2025); Starnini et al. (2025); Sastre & Henkel (2016); Dornic et al. (2001), the study of LLM calibration and intrinsic biases Santagata & De Nobili (2024); Zheng et al. (2024); Germani & Spitale (2025); Nadeem et al. (2026); Knipper et al. (2025), simulating opinion dynamics with LLM agents and emergent cultural attractors Chuang et al. (2024); Perez et al. (2025); Johnson (2026), mean-field treatments of multi-agent learning and behavior modeling Yang et al. (2018); Cozzi et al. (2025), agents econ-

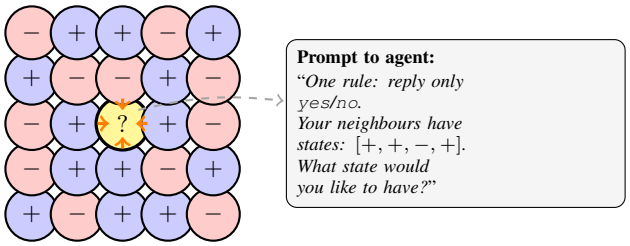


Figure 1: Schematic of the LLM-on-lattice setup. Each site of the $L \times L$ square lattice (periodic boundary conditions) hosts an identical LLM agent in a binary state $s_i \in \{+1, -1\}$, mapped to `yes`/`no` (blue for $+$, red for $-$). At each micro-update, a site is chosen at random (yellow, central) and its agent is queried with the prompt shown on the right, conditioned on its four nearest-neighbor states (orange arrows). The reply is parsed and written back as s_i'. One *sweep* consists of $N = L^2$ such micro-updates.

omy Tomasev et al. (2025); Wang et al. (2026), decentralized AI Sun et al. (2023); Ding et al. (2025); Riehl et al. (2026), and recent investigations into the equilibrium structure of LLM-driven dynamics Song et al. (2025).

2 Model and Methods

2.1 Lattice, states, and update rule

We consider an $L \times L$ square lattice with periodic boundary conditions, $N = L^2$ sites. Each site i holds a binary state $s_i \in \{-1, +1\}$, mapped to textual labels (`no`, `yes`). Updates follow an asynchronous Monte Carlo-like protocol: at each micro-step a site (i, j) is chosen uniformly at random and its LLM agent is queried with a minimal prompt of the form

```
One rule: reply only yes/no. Your neighbours have states:
[...]. What state would you like to have?
```

The reply is parsed to $s_i' \in \{-1, +1\}$ and written back. One *sweep* consists of N such updates. The LLM sampler temperature T is the sole external control parameter. Figure 1 provides a schematic illustration of the setup. For reference, the classical 2D Ising Hamiltonian is

$$H = -J \sum_{\langle i,j \rangle} s_i s_j - h \sum_i s_i \,, \tag{1}$$

where J is the nearest-neighbor coupling and h an external field. Onsager's exact solution at $h = 0$ gives a continuous phase transition at $\beta_c J = \frac{1}{2} \ln(1 + \sqrt{2}) \approx 0.4407$, with universal critical exponents $\gamma/\nu = 7/4$ in two dimensions.

2.2 Global-flip protocol

LLMs generically prefer one label over the other (e.g., `yes` over `no`), introducing a nonzero effective field that breaks $\mathbb{Z}_2$ symmetry. To probe and partially cancel this bias, we run paired ensembles with a global relabeling $s_i \to -s_i$ (equivalently swapping `yes`↔`no` in all prompts and responses; see Appendix A). Averaging even moments ($\langle m^2 \rangle$, $\langle m^4 \rangle$) across both flip sectors cancels odd-in-m bias in expectation, while odd moments and the absolute magnetization $\langle |m| \rangle$ can be compared across sectors to quantify the residual field.

We verify the protocol directly by comparing the late-time magnetization distribution $P(m)$ across sectors (Appendix A, Fig. 5): `phi4-mini:3.8b` and `llama3.1:8b` produce approximately mirror-symmetric distributions, validating the pooling procedure, whereas `mistral:7b` shows visibly broken symmetry, an early signature of the persistently strong effective field analyzed in Sec. 3.3.

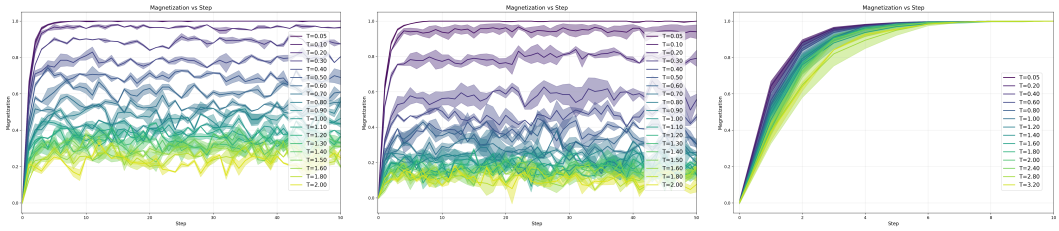


Figure 2: Magnetization $m(t)$ versus sweep number at various sampler temperatures for `llama3.1:8b` (a), `phi4-mini:3.8b` (b), and `mistral:7b` (c). Shaded bands denote $\pm 1\sigma$ across seeds. `llama3.1:8b` and `phi4-mini:3.8b` show clear temperature-driven disordering, while `mistral:7b` reaches near-full alignment at all temperatures, a direct manifestation of its persistently strong effective field (Fig. 4).

2.3 Observables

From the magnetization time series $m(t) = N^{-1} \sum_i s_i(t)$ after a burn-in window, we estimate late-time averages $\langle m^k \rangle$ by averaging over sweeps and then over seeds. We compute two susceptibility variants,

$$\chi = \frac{N}{T}\big(\langle m^2 \rangle - \langle m \rangle^2\big), \quad \chi_{|m|} = \frac{N}{T}\big(\langle m^2 \rangle - \langle |m| \rangle^2\big). \tag{2}$$

When both flip sectors are pooled, $\langle m \rangle \approx 0$ by construction, so χ reduces to $N\langle m^2 \rangle / T$ and diverges trivially as $1/T$ at low temperature. The absolute susceptibility $\chi_{|m|}$ avoids this artifact by subtracting $\langle |m| \rangle^2$, which correctly captures the magnitude of alignment in both sectors; we therefore use $\chi_{|m|}$ throughout. In finite-size scaling, a genuine 2D Ising transition predicts $\chi_{\max}(L) \sim L^{\gamma/\nu}$ with $\gamma/\nu = 7/4 = 1.75$.

2.4 Effective coupling and field from update logs

For each micro-update we record the local neighbor field $k = \sum_{j \in \partial i} s_j \in \{-4, -2, 0, 2, 4\}$ and the resulting spin s_i'. Aggregating over all updates at a given T, we fit the log-odds to a linear model,

$$\text{logit}\ P(s_i' = +1 \mid k) \ \approx\ 2(\tilde{h} + \tilde{J}\, k)\,, \tag{3}$$

where $\tilde{J} = \beta J$ and $\tilde{h} = \beta h$ are effective β-weighted parameters. This logistic form is exact for a heat-bath update of the Ising Hamiltonian, Eq. (1); applied to LLM updates, it provides a model-agnostic decomposition into coupling strength ("social conformity") and bias ("intrinsic preference") that requires no access to the model's internal weights or architecture. We apply Jeffreys smoothing to avoid infinite logits in low-count bins (see Appendix B). For models with strong bias at low temperature (notably `mistral:7b`), certain local-field bins become extremely rare; in these cases we supplement the standard binning with a rare-event analysis that pools statistics across neighboring temperature points to stabilize the fit (Appendix B).

2.5 Simulation parameters

We study three open-weight models (`llama3.1:8b`, `phi4-mini:3.8b`, and `mistral:7b`) served locally via Ollama Ollama (2024). We simulate $L = 8, 10, 12, 16, 20, 24$ for `llama3.1:8b`, $L = 10, 12, 16, 20, 24$ for `phi4-mini:3.8b`, and $L = 5, 10, 15, 20, 25, 30, 35$ for `mistral:7b`. Finite-size scaling is restricted to even-L lattices, for the reasons discussed in Appendix C: every simulated size is even for `llama3.1:8b` and `phi4-mini:3.8b`, so all of them enter the γ/ν fit, whereas for `mistral:7b` only $L = 10, 20, 30$ do. This is why the `mistral:7b` susceptibility curves are shown up to $L = 35$ while γ/ν is computed up to $L = 30$. Sampler temperatures are scanned from $T = 0.05$ to $T \approx 3.2$. For each (L, T) we run at least six independent seeds, equally split between $g = +1$ and $g = -1$ flip sectors. The total number of sweeps per run scales as $2L^2$ near the crossover temperature, ensuring that the burn-in window (70% of total sweeps) comfortably exceeds the system's relaxation time, which grows as $\sim L^2$ for Glauber-like dynamics. Observables are computed over the remaining 30% measurement window,

and uncertainties are estimated by bootstrap over seeds. All plots show $\pm 1\sigma$ error bands from the seed-averaging procedure. All production runs were executed on identical hardware (NVIDIA RTX 5090) to eliminate architecture-dependent variability in LLM inference (see Appendix C).

3 RESULTS

3.1 MAGNETIZATION DYNAMICS

Figure 2 shows the magnetization time series $m(t)$ for all three models across sampler temperatures. The qualitative behavior is shared: at low T, the system rapidly reaches high alignment; as T increases, trajectories become noisier and the steady-state magnetization drops. However, the rate and extent of this disordering differ markedly across models.

`llama3.1:8b` [Fig. 2(a)] reaches near-full alignment ($m \to 1$) at $T \leq 0.2$ and shows a gradual decrease with clear temperature separation. At intermediate temperatures ($T \approx 0.4$–0.8), trajectories are noisy with substantial inter-seed variance, while at $T \geq 1.0$ the magnetization fluctuates around a reduced but still positive value ($m \approx 0.2$–0.3), reflecting the residual bias.

`phi4-mini:3.8b` [Fig. 2(b)] shows a similar pattern but compressed in temperature: the transition from high to low magnetization occurs over a narrower T range, consistent with the steep decay of this model's effective field (Sec. 3.3).

`mistral:7b` [Fig. 2(c)] is strikingly different. The system reaches near-complete alignment ($m \to 1$) at *all* sampled temperatures, including $T = 3.2$. This behavior is a direct signature of the persistently strong effective field for this model (Sec. 3.3): even at high sampler temperatures, the intrinsic label preference dominates over stochastic noise.

3.2 SUSCEPTIBILITY AND FINITE-SIZE SCALING

Figure 3 shows the absolute susceptibility $\chi_{|m|}(T)$ for all three models on even-L lattices. Every model exhibits a susceptibility peak, indicating a temperature of maximal consensus fragility. However, the character of the peaks and the resulting finite-size scaling differ qualitatively, yielding three distinct effective exponents extracted from weighted log-log fits of $\chi_{\max}(L)$:

$$\gamma/\nu = 1.02 \pm 0.05 \quad (\texttt{llama3.1:8b}),$$
$$\gamma/\nu = 1.75 \pm 0.13 \quad (\texttt{phi4-mini:3.8b}),$$
$$\gamma/\nu = 2.01 \pm 0.04 \quad (\texttt{mistral:7b}).$$

The fact that these exponents are all different is itself strong evidence against universal Ising behavior: in a genuine universality class, the exponent would be the same regardless of microscopic details.

`phi4-mini:3.8b` [Fig. 3(b)] displays sharp, well-defined peaks centered at $T \approx 0.2$, with peak height growing cleanly with L. The effective exponent $\gamma/\nu = 1.75 \pm 0.13$ is statistically compatible with the 2D Ising value of $7/4 = 1.75$. However, this proximity is not evidence for genuine Ising criticality. As shown in Sec. 3.3, $\tilde{J} \approx 0$ for `phi4-mini:3.8b` at all temperatures, meaning there is essentially no neighbor coupling. A diverging correlation length, the mechanism that produces $\gamma/\nu = 7/4$ in the Ising model, requires finite coupling. A system with vanishing $\tilde{J}$ cannot develop the long-range correlated domains characteristic of a critical point. We therefore interpret the proximity to $7/4$ as a coincidence that reflects the steepness of `phi4-mini:3.8b`'s field crossover: $\tilde{h}(T)$ drops from ~ 2.7 to near zero in a narrow window $\Delta T \approx 0.2$, producing a sharp peak in $\chi_{|m|}$ whose L-scaling, over the limited range of accessible system sizes, happens to mimic Ising-like behavior.

`llama3.1:8b` [Fig. 3(a)] shows intermediate behavior, with peaks around $T \approx 0.4$–0.5. The effective exponent $\gamma/\nu = 1.02 \pm 0.05$ corresponds to $\chi_{\max} \sim L$ (or equivalently $\sqrt{N}$), a sub-linear growth that lies between the regime of independent agents ($\gamma/\nu = 0$) and the volume-scaling regime ($\gamma/\nu = 2$). Physically, this indicates weak but real cooperative amplification: `llama3.1:8b` agents are not independent, but their correlations do not extend to the long ranges characteristic of a critical system.

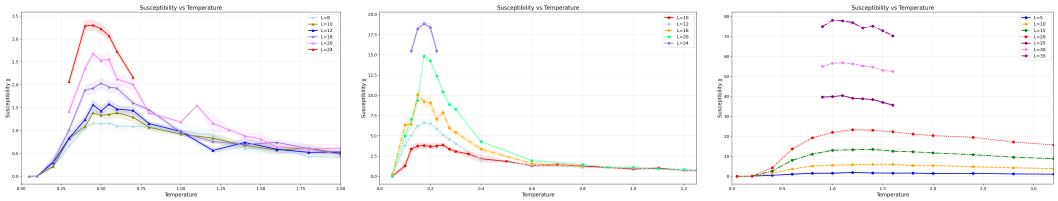


Figure 3: Absolute susceptibility $\chi_{|m|}(T)$ on even-L lattices for (a) `llama3.1:8b`, (b) `phi4-mini:3.8b`, and (c) `mistral:7b`. Simulated sizes are $L = 8, 10, 12, 16, 20, 24$, $L = 10, 12, 16, 20, 24$ and $L = 5, 10, 15, 20, 25, 30, 35$ respectively; γ/ν is fitted on even L only, i.e. $L = 10, 20, 30$ for `mistral:7b`. Shaded bands show $\pm 1\sigma$ seed uncertainty. (a) Peaks around $T \approx 0.4$–0.5 give $\gamma/\nu = 1.02 \pm 0.05$, i.e. $\chi_{\max} \sim L$, a sub-linear growth indicating weak collective amplification. (b) Sharp peaks at $T \approx 0.2$ give $\gamma/\nu = 1.75 \pm 0.13$: the central value is statistically compatible with the 2D Ising exponent, but the vanishing effective coupling $\tilde{J} \approx 0$ (Fig. 4) rules out genuine Ising criticality, and we interpret the proximity as a coincidence reflecting the steepness of the field crossover (see text). (c) Very broad peaks around $T \approx 1.0$–1.2 grow as $\gamma/\nu = 2.01 \pm 0.04$, corresponding to trivial volume scaling $\chi_{\max} \propto N$: a field-dominated system whose fluctuations scale with size rather than with a diverging correlation length.

`mistral:7b` [Fig. 3(c)] produces the largest susceptibility values ($\chi \sim 80$ at $L = 35$) with very broad peaks centered around $T \approx 1.0$–1.2 that grow rapidly with system size. The effective exponent $\gamma/\nu = 2.01 \pm 0.04$ saturates the dimensional ceiling $d = 2$ and corresponds exactly to volume scaling $\chi_{\max} \propto L^2 = N$. This is the expected behavior for a uniformly magnetized system with small Gaussian fluctuations: the variance of the total magnetization scales as N by the central limit theorem, and dividing by T gives $\chi \propto N/T$. The persistent $\tilde{h}$ for `mistral:7b` (Sec. 3.3) confirms that the system never truly disorders; the broad "peak" marks where the field weakens enough for volume-scaled fluctuations to become visible, not a critical point.

3.3 Effective couplings and fields

Figure 4 presents the central diagnostic of this study: the effective coupling $\tilde{J}(T)$ and effective field $\tilde{h}(T)$ extracted from update logs for all three models. The horizontal dashed line marks the 2D Ising critical coupling $\beta_c J \approx 0.4407$.

Across all models and temperatures, $\tilde{h} \gg \tilde{J}$: the field dominates while the coupling remains near zero. This single observation explains both the absence of Ising criticality and the model-dependent effective exponents reported in Sec. 3.2. A second-order phase transition in the Ising universality class requires $h = 0$ and finite J; when the field dominates, the transition is replaced by a smooth crossover whose steepness depends on the rate at which $\tilde{h}(T)$ decays.

The three models trace qualitatively distinct trajectories in $(\tilde{J}, \tilde{h})$ space, and these trajectories map directly onto the susceptibility signatures. `phi4-mini:3.8b` (green) exhibits the steepest field decay: $\tilde{h}$ drops from ~ 2.7 to near zero by $T \approx 0.3$, with $\tilde{J} \approx 0$ throughout. This steep crossover produces the sharp susceptibility peak and the exponent $\gamma/\nu = 1.75 \pm 0.13$ of Fig. 3(b). `llama3.1:8b` (blue) shows intermediate field decay, with $\tilde{J} \approx 0.35$ at the lowest temperature (the only model where the coupling approaches the Ising critical value), but $\tilde{h} \approx 2.8$ at the same point, so the field still dominates by nearly an order of magnitude. `mistral:7b` (red) has the slowest field decay: $\tilde{h}$ starts at ~ 1.3 (at $T = 0.4$) and remains substantial across the full sampled range, while its coupling $\tilde{J}$ fluctuates around 0.1–0.2 with large uncertainty. This persistent field explains the broad susceptibility plateau and the volume scaling $\gamma/\nu = 2.01 \pm 0.04$ of Fig. 3(c). The physical interpretation is sharp: most of the time, what looks like collective consensus at low temperature is not the result of agents influencing each other. It is the result of all agents independently following the same intrinsic bias. This is "consensus by shared preference," not "consensus by deliberation."

Taken together, these results rank the three models in terms of genuine social behavior: `llama3.1:8b` is the most social, with sub-linear scaling reflecting real (if modest) inter-agent influence; `phi4-mini:3.8b` is essentially asocial, with vanishing $\tilde{J}$ producing an Ising-like ex-

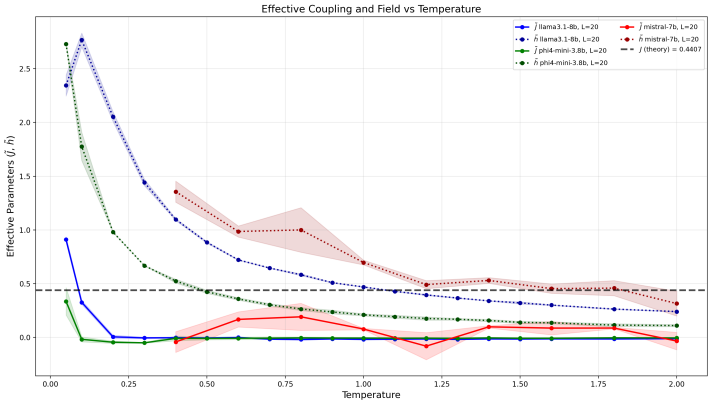


Figure 4: Effective coupling $\tilde{J}$ (solid lines) and field $\tilde{h}$ (dotted lines) versus sampler temperature for `llama3.1:8b` (blue), `phi4-mini:3.8b` (green), and `mistral:7b` (red), measured at $L = 20$. Shaded bands show $\pm 1\sigma$ seed uncertainty. The horizontal dashed line marks the 2D Ising critical coupling $\beta_c J \approx 0.4407$. For all models, $\tilde{h} \gg \tilde{J}$: collective alignment is field-dominated rather than coupling-driven. The rate at which $\tilde{h}(T)$ decays (steep for `phi4-mini:3.8b`, moderate for `llama3.1:8b`, slow for `mistral:7b`) directly determines the character of the susceptibility peak and the effective γ/ν (Fig. 3).

ponent that has nothing to do with cooperation; and `mistral:7b` is the least social, with alignment driven almost entirely by its persistent field.

4 Discussion

4.1 Why deviations from Ising universality are expected

The 2D Ising universality class rests on three conditions: a $\mathbb{Z}_2$-symmetric Hamiltonian ($h = 0$), detailed balance (convergence to a Gibbs measure), and short-range interactions. Our LLM lattice generically violates the first two.

The broken $\mathbb{Z}_2$ symmetry is manifest in the large $\tilde{h}$ values of Fig. 4 and in the asymmetric $P(m)$ distributions of Fig. 5: the LLM's weights and alignment training encode a systematic label preference that acts as an external field. The global-flip protocol cancels this bias at the level of averaged even moments, but does not restore microscopic $\mathbb{Z}_2$ symmetry. Each individual trajectory still evolves under the biased update kernel, so within-run correlations carry the field's imprint. In equilibrium Ising with $h \neq 0$, there is no phase transition at all; the singularity is replaced by a smooth crossover, which is precisely what we observe.

Detailed balance is the second concern. The LLM update kernel is an empirical object: while the logistic fit, Eq. (3), implies consistency with *some* effective Hamiltonian when the fit is good, there is no guarantee that the actual conditional probabilities satisfy detailed balance with respect to any energy function. If they do not, the stationary distribution (if one exists) is not a Gibbs state, and equilibrium critical exponents need not apply. Independent evidence for non-equilibrium dynamics in our setup is provided by the lattice-parity effect described in Appendix C: in an equilibrium system, $\chi_{\max}(L)$ is monotonically non-decreasing regardless of parity, whereas persistent probability currents in configuration space can couple to the lattice topology in L-dependent ways.

Finally, the LLM's decision is not strictly a function of the scalar local field $k = \sum_{j \in \partial i} s_j$. The agent processes a token sequence and may be sensitive to the spatial arrangement of neighbors, prompt syntax, and token ordering. If the effective update rule depends on more than k, the system departs from the Ising universality class even when other conditions are met.

4.2 THE SAMPLER TEMPERATURE AS A CONTROL PARAMETER

The sampler temperature rescales logits before the softmax, providing a well-defined noise axis: $T \to 0$ gives deterministic decoding and $T \to \infty$ gives uniform sampling. However, increasing T simultaneously reduces both $\tilde{J}$ and $\tilde{h}$, at rates that differ across models (Fig. 4). The system therefore moves along a *curve* in $(\tilde{J}, \tilde{h})$ space rather than along the $h = 0$ line where the Ising transition lives. This explains why each model traces a different crossover trajectory and why no universal T_c emerges.

4.3 THE SPECTRUM OF EFFECTIVE EXPONENTS

The three measured exponents ($\gamma/\nu \approx 1.02$ for `llama3.1:8b`, ≈ 1.75 for `phi4-mini:3.8b`, ≈ 2.01 for `mistral:7b`) are not random scatter but correlate systematically with the rate at which $\tilde{h}(T)$ decays. A steep field decay (`phi4-mini:3.8b`) concentrates the crossover in a narrow T window, producing a sharp susceptibility peak whose L-scaling happens to match the Ising value within uncertainty, despite the absence of any genuine critical mechanism. A slow decay (`mistral:7b`) keeps the system magnetized at all accessible temperatures, so $\chi_{\max}$ scales trivially with volume (L^2). An intermediate decay (`llama3.1:8b`) yields a near-linear growth in L, reflecting weak collective amplification diluted over a broad crossover. This pattern suggests a unified interpretation: the effective exponent is not a critical exponent in the renormalization-group sense, but rather a measure of how abruptly the LLM's intrinsic bias switches off as stochasticity increases.

4.4 IMPLICATIONS FOR MULTI-AGENT SYSTEMS

Independent of whether strict Ising universality holds, the $(\tilde{J}, \tilde{h})$ decomposition provides actionable diagnostics for LLM multi-agent design. A large $\tilde{h}/\tilde{J}$ ratio signals that apparent consensus is driven by shared bias rather than genuine collective reasoning, a form of "echo chamber by default." The susceptibility peak locates the temperature of maximal consensus fragility, where small perturbations can most effectively shift group opinion.

The framework also reveals intrinsic properties of the LLM models themselves. `mistral:7b` acts as a strongly biased agent that maintains alignment over a wide temperature range, reflecting a model whose token preferences are deeply embedded and resistant to contextual override. `phi4-mini:3.8b` transitions sharply from biased order to independent disorder, suggesting a model whose bias is strong but narrowly concentrated in logit space and easily disrupted by temperature. `llama3.1:8b` occupies an intermediate regime, showing the largest coupling of the three models (though still modest in absolute terms) and the most pronounced non-equilibrium signatures. These collective-behavior fingerprints expose aspects of each model's decision-making that would be difficult to access through standard single-agent evaluations, and they suggest a new class of benchmarks for comparing LLMs as social agents.

4.5 THE $\tilde{h}/\tilde{J}$ RATIO AS A CONSENSUS RELIABILITY MEASURE

The ratio $\tilde{h}/\tilde{J}$ has a direct practical interpretation: it measures how much of observed collective alignment is attributable to shared bias versus cooperative interaction. When $\tilde{h}/\tilde{J} \gg 1$, as in all three models studied here, the multi-agent consensus is effectively a single-agent opinion amplified N times; the collective provides no additional epistemic value over a single query. Only when $\tilde{h}/\tilde{J} \lesssim 1$ does agreement among agents carry genuine evidential weight, because it reflects inter-agent influence rather than correlated bias.

This diagnostic can be applied before deploying any MAS pipeline. For instance, `llama3.1:8b` is the only model in our study with measurable coupling ($\tilde{J} \approx 0.35$ at low T), making it the strongest candidate for tasks requiring genuine deliberation, such as iterative debate or collaborative reasoning. However, even for `llama3.1:8b`, $\tilde{h}/\tilde{J} \approx 8$ at the lowest temperature, meaning bias still dominates by nearly an order of magnitude. This sobering finding suggests that, for current open-weight models under minimal prompting, multi-agent consensus should be treated as "amplified single-agent opinion" rather than "deliberated group judgment." Achieving the cooperative regime

($\tilde{h}/\tilde{J} \lesssim 1$) likely requires targeted prompt engineering or architectural modifications, and the $(\tilde{J}, \tilde{h})$ framework provides the quantitative feedback loop needed to guide such efforts.

5 Limitations and Outlook

Several limitations constrain our current results. Lattice sizes ($L \leq 35$) and sweep counts are limited by the cost of LLM inference, restricting finite-size scaling to roughly half a decade in L. Burn-in and equilibration are difficult to certify, and strong prompt dependence means that results may shift with alternative phrasings. The asynchronous update protocol introduces additional non-equilibrium effects.

These limitations motivate a clear roadmap for future work. First, prompt design could be optimized to minimize the effective field ("field tuning"), potentially enabling access to the $h \approx 0$ regime where a genuine transition might appear. Preliminary experiments with modified prompts show that prompt changes can dramatically alter the $(\tilde{J}, \tilde{h})$ balance and can break the $\mathbb{Z}_2$ symmetry of the global-flip protocol, highlighting both the promise and the delicacy of this approach. Second, direct tests of detailed balance (checking whether the empirical transition rates are consistent with *any* energy function) would determine whether an equilibrium description is even approximately valid. Third, comparisons against known non-equilibrium models (voter dynamics, majority rule) on the same lattice would help classify the dynamics when equilibrium descriptions fail. Fourth, larger lattices, longer runs, and autocorrelation-time measurements would sharpen finite-size scaling and test whether the near-Ising exponent of `phi4-mini:3.8b` persists at larger L. Finally, extensions to richer state spaces, longer-range interactions, and heterogeneous agents could connect this framework to more realistic multi-agent coordination scenarios.

6 Conclusion

We have introduced a minimal, model-agnostic statistical-physics framework for characterizing collective behavior in 2D lattices of LLM agents with local binary interactions. It requires no access to the model's internal weights or architecture, operating entirely from observable update logs, and it allows one to disentangle intrinsic bias from cooperation, compute critical exponents, and diagnose crossovers and possible phase transitions in arbitrary multi-agent systems. Across three open-weight models we find order-disorder crossovers driven by sampler temperature, susceptibility peaks marking regions of maximal consensus fragility, and a systematic dominance of the effective field over the effective coupling ($\tilde{h} \gg \tilde{J}$), which replaces the Ising transition with a model-dependent crossover and yields a spectrum of non-universal exponents correlated with the rate at which $\tilde{h}(T)$ decays.

Rather than a negative result, this is a quantitative characterization of how LLM collectives reach apparent consensus: low-temperature alignment is "consensus by shared preference" rather than "consensus by deliberation." The practical consequence for multi-agent alignment is concerning. Architectures that rely on agreement among copies of the same model (debate, self-consistency, committee voting) cannot assume that N-agent consensus is more informative than a single query when $\tilde{h} \gg \tilde{J}$, since the agreement is correlated through shared bias rather than produced by independent reasoning; the concern is especially acute in decentralized settings, where no central authority can correct for shared biases. Together with the susceptibility peak, which locates the temperature at which a collective is most responsive to external steering, the $(\tilde{J}, \tilde{h})$ decomposition forms a compact toolkit for evaluating and comparing LLM multi-agent systems, and a starting point for a statistical-physics research program on prompt-based field tuning toward the cooperative regime ($\tilde{h}/\tilde{J} \lesssim 1$), detailed-balance diagnostics, and non-equilibrium universality in collective artificial intelligence.

Reproducibility Statement

All simulation parameters needed to reproduce our results are specified in Sec. 2: the lattice geometry and update rule (Sec. 2.1), the exact prompt, the global-flip protocol (Sec. 2.2 and Appendix A), the observables and susceptibility definitions (Sec. 2.3), the logistic-regression estimator for $(\tilde{J}, \tilde{h})$ including Jeffreys smoothing and the rare-event pooling (Sec. 2.4 and Appendix B), and the lattice

sizes, temperature grid, seed counts, burn-in fraction and bootstrap procedure (Sec. 2.5 and Appendix C). Appendix C additionally documents the even-L restriction and its rationale, the weighted fit used to extract γ/ν, and the hardware constraint imposed to avoid architecture-dependent drift in LLM inference. All models are open-weight and served through a standard local runtime, so no proprietary API access is required to reproduce the study.

USE OF AI ASSISTANCE

In accordance with the ICLR 2027 policy on the use of AI tools, we disclose the following. An AI assistant (a large language model) was used for two purposes in the preparation of this manuscript: (i) retrieval and discovery, to help identify related work in the alignment literature (specifically the references on learning from human preferences, instruction tuning, sycophancy, and LLM-as-judge evaluation cited in Sec. 1); and (ii) assistance in drafting and polishing the framing of the introduction connecting the $(\tilde{J}, \tilde{h})$ decomposition to that literature. All such text was reviewed, edited, and verified by the author, and all cited references were checked against their primary sources. No AI tool was used to design the experiments, run the simulations, analyze the data, produce the figures, or derive any of the quantitative results reported in this work. The author takes full responsibility for the entire content of the paper.

REFERENCES

Claudio Castellano, Santo Fortunato, and Vittorio Loreto. Statistical physics of social dynamics. *Rev. Mod. Phys.*, 81:591–646, 2009. doi: 10.1103/RevModPhys.81.591.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers. Simulating opinion dynamics with networks of LLM-based agents, 2024. URL https://arxiv.org/abs/2311.09618.

Francesco Cozzi, Marco Pangallo, Alan Perotti, André Panisson, and Corrado Monti. Learning individual behavior in agent-based models with graph diffusion networks, 2025. URL https://arxiv.org/abs/2505.21426. Adv. Neural Inf. Process. Syst. 38.

Yepeng Ding, Ahmed Twabi, Junwei Yu, Lingfeng Zhang, Tohru Kondo, and Hiroyuki Sato. Decentralized multi-agent system with trust-aware communication. In *2025 IEEE International Symposium on Parallel and Distributed Processing with Applications (ISPA)*, pp. 1439–1445. IEEE, Oct 2025. doi: 10.1109/ispa67752.2025.00198.

Ivan Dornic, Hugues Chaté, Jérôme Chave, and Haye Hinrichsen. Critical coarsening without surface tension: the voter universality class. *Phys. Rev. Lett.*, 87:045701, 2001. doi: 10.1103/PhysRevLett.87.045701.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate, 2023. URL https://arxiv.org/abs/2305.14325.

Federico Germani and Giovanni Spitale. Source framing triggers systematic evaluation bias in large language models, 2025. URL https://arxiv.org/abs/2505.13488.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges, 2024. URL https://arxiv.org/abs/2402.01680.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework, 2024. URL https://arxiv.org/abs/2308.00352.

Neil F. Johnson. Increasing intelligence in AI agents can worsen collective outcomes, 2026. URL https://arxiv.org/abs/2603.12129.

R. Alexander Knipper, Charles S. Knipper, Kaiqi Zhang, Valerie Sims, Clint Bowers, and Santu Karmaker. The bias is in the details: An assessment of cognitive bias in llms, 2025. URL https://arxiv.org/abs/2509.22856.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. A survey on LLM-as-a-Judge, 2024. URL https://arxiv.org/abs/2411.15594.

Pratik Mullick and Parongama Sen. Sociophysics models inspired by the Ising model. *Eur. Phys. J. B*, 98:206, 2025. doi: 10.1140/epjb/s10051-025-01053-7.

Afrozah Nadeem, Usman Naseem, Yuan-Fang Ge, and Imran Razzak. Bias beyond borders: Political ideology evaluation and steering in multilingual LLMs, 2026. URL https://arxiv.org/abs/2601.23001.

Ollama. Ollama: Run large language models locally. https://ollama.com, 2024.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022. URL https://arxiv.org/abs/2203.02155.

Jeremy Perez, Grgur Kovač, Cédric Léger, Cédric Colas, Gaia Molinaro, Maxime Derex, Pierre-Yves Oudeyer, and Clément Moulin-Frier. When LLMs play the telephone game: cultural attractors as conceptual tools to evaluate LLMs in multi-turn settings, 2025. URL https://arxiv.org/abs/2407.04503. Proc. ICLR 2025.

Giorgio Piatti, Zhijing Hu, and Kyunghyun Cho. Cooperate or collapse: Emergence of sustainable cooperation in a society of LLM agents, 2024. URL https://arxiv.org/abs/2404.16698.

Kevin Riehl, Julius Schlapbach, Anastasios Kouvelas, and Michail A. Makridis. Karma mechanisms for decentralised, cooperative multi agent path finding, 2026. URL https://arxiv.org/abs/2604.07970.

Chenxi Ruan, Yiming Wang, Zhenyu Shi, André Panisson, and Jiaxin Li. Reaching agreement among reasoning LLM agents, 2025. URL https://arxiv.org/abs/2512.20184.

Luca Santagata and Cristiano De Nobili. More is more: Addition bias in large language models, 2024. URL https://arxiv.org/abs/2409.02569.

Francisco Sastre and Malte Henkel. Antiferromagnetic majority voter model on square and honeycomb lattices. *Physica A*, 444:897–905, 2016. doi: 10.1016/j.physa.2015.10.083.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023. URL https://arxiv.org/abs/2310.13548.

Andries Smit, Paul Duckworth, Nathan Grinsztajn, Thomas D. Barrett, and Arnu Pretorius. Should we be going MAD? a look at multi-agent debate strategies for LLMs, 2024. URL https://arxiv.org/abs/2311.17371.

Zi-Yu Song, Qing-Hong Cao, Ming-Xing Luo, and Hua Xing Zhu. Detailed balance in large language model-driven agents, 2025. URL https://arxiv.org/abs/2512.10047.

Michele Starnini, Fabian Baumann, Tobias Galla, David Garcia, Gerardo Iñiguez, Márton Karsai, Jan Lorenz, and Katarzyna Sznajd-Weron. Opinion dynamics: Statistical physics and beyond, 2025. URL https://arxiv.org/abs/2507.11521.

Xinyuan Sun, Davide Crapis, Matt Stephenson, Barnabé Monnot, Thomas Thiery, and Jonathan Passerat-Palmbach. Cooperative AI via decentralized commitment devices, 2023. URL https://arxiv.org/abs/2311.07815.

Nenad Tomasev, Matija Franklin, Joel Z. Leibo, Julian Jacobs, William A. Cunningham, Iason Gabriel, and Simon Osindero. Virtual agent economies, 2025. URL https://arxiv.org/abs/2509.10147.

Shouqiao Wang, Marcello Politi, Samuele Marro, and Davide Crapis. Profit is the red team: Stress-testing agents in strategic economic interactions, 2026. URL https://arxiv.org/abs/2603.20925.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models, 2022. URL https://arxiv.org/abs/2203.11171. Published at ICLR 2023.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80 of *PMLR*, pp. 5571–5580, 2018.

Haitao Zheng, Xiaoyan Liu, Lingyu Wang, Xinyu Yang, Yifan Liu, and Xianpei Wang. CalibraEval: Calibrating prediction distribution to mitigate selection bias in LLMs-as-judges, 2024. URL https://arxiv.org/abs/2410.15393.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2023.

A GLOBAL-FLIP MAPPING DETAILS

For each simulation run, a global parity $g \in \{-1, +1\}$ is drawn. When $g = +1$ nothing happens. When $g = -1$, all prompts swap `yes`↔`no` in both the neighbor-state description and the response parsing: an LLM reply of `yes` is recorded as $s_i = -1$, and vice versa. Observables that are even in m ($\langle m^2 \rangle$, $\langle m^4 \rangle$) are invariant under this relabeling by construction. Averaging odd quantities ($\langle m \rangle$, $\langle |m| \rangle$) across $g = \pm 1$ sectors reveals the magnitude of the effective field: a large difference $\langle m \rangle_{g=+1} - \langle m \rangle_{g=-1}$ signals strong label bias.

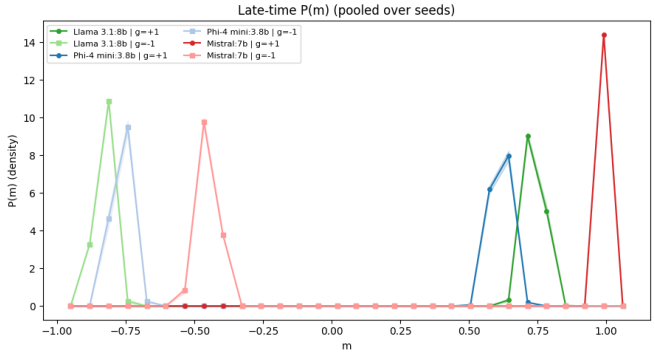


Figure 5: Late-time magnetization distribution $P(m)$ pooled over seeds, separated by flip sector ($g = +1$, $g = -1$), for all three models at their respective susceptibility-peak temperatures ($T = 0.225$ for `phi4-mini:3.8b`, $T = 0.45$ for `llama3.1:8b`, $T = 1.2$ for `mistral:7b`; all at $L = 20$). `phi4-mini:3.8b` and `llama3.1:8b` show approximate mirror symmetry between sectors, validating the global-flip protocol. `mistral:7b` shows strongly broken symmetry: the $g = +1$ sector locks at $m \approx 1$ while the $g = -1$ sector reaches only $m \approx -0.5$, reflecting its persistent effective field.

Figure 5 shows $P(m)$ pooled over seeds and separated by flip sector, for all three models at their respective susceptibility-peak temperatures. For `phi4-mini:3.8b` ($T = 0.225$) and `llama3.1:8b` ($T = 0.45$), the two sectors produce approximately mirror-symmetric distributions centered at $m \approx \pm 0.65$ and $m \approx \pm 0.8$ respectively, with zero overlap between sectors; this approximate symmetry validates the pooling procedure for even-moment observables. For `mistral:7b` ($T = 1.2$) the symmetry is visibly broken: the $g = +1$ sector concentrates at $m \approx +1.0$ (near-complete alignment) while the $g = -1$ sector peaks at $m \approx -0.5$ (partial alignment). This asymmetry reflects `mistral:7b`'s persistently strong effective field (Sec. 3.3), which prevents the global-flip protocol from fully equalizing the two sectors.

B Logistic regression and rare-event analysis details

For each (T, model), we aggregate all micro-updates across seeds and lattice sweeps. Updates are binned by the local field $k \in \{-4, -2, 0, 2, 4\}$, and the empirical probability $\hat{p}_+(k)$ is computed with Jeffreys smoothing: $\hat{p}_+(k) = (n_+(k) + \frac{1}{2})/(n(k) + 1)$, where $n_+(k)$ is the count of $s'_i = +1$ outcomes and $n(k)$ the total count at field value k. A weighted linear regression of $\operatorname{logit} \hat{p}_+(k)$ on k yields $\tilde{h}$ (intercept divided by 2) and $\tilde{J}$ (slope divided by 2). Weights are proportional to $n(k)$, ensuring that well-sampled field values dominate the fit. Bins with $n(k) < 5$ are excluded as a robustness check; results are stable to this threshold.

For `mistral:7b`, the strong effective field at low and intermediate temperatures means that local-field bins opposing the preferred label (e.g., $k = -4$ when the bias favors $+1$) are populated by very few updates. Standard binning yields unreliable logit estimates in these rare-event bins. We address this by pooling update counts across a sliding window of neighboring temperature points, providing sufficient statistics for stable logistic regression while preserving the temperature dependence of $\tilde{J}$ and $\tilde{h}$. The larger error bands for `mistral:7b` in Fig. 4 reflect the residual uncertainty from this procedure.

C Technical details and reproducibility

Even-L lattices and the parity effect. All susceptibility scaling analyses presented in this work use even-L lattices: all simulated sizes are even for `llama3.1:8b` ($L = 8, 10, 12, 16, 20, 24$) and `phi4-mini:3.8b` ($L = 10, 12, 16, 20, 24$), while for `mistral:7b` the even subset is $L = 10, 20, 30$. This choice is motivated by the fact that even-L lattices with periodic boundary conditions are bipartite, which preserves the antiferromagnetic symmetry of the underlying Hamiltonian and produces cleaner finite-size scaling. Odd-L lattices, by contrast, contain frustrated loops induced by the periodic boundary conditions and can introduce subleading topological corrections that distort scaling fits at small system sizes. In our exploratory runs we observed a systematic lattice-parity effect: odd-L and even-L subsequences yielded different effective exponents, with the largest discrepancy observed for `llama3.1:8b`. We interpret this parity effect as a finite-size artifact amplified by the non-equilibrium character of the LLM update kernel: in an equilibrium system, $\chi_{\max}(L)$ is monotonically non-decreasing regardless of parity, but in a non-equilibrium system the persistent probability currents in configuration space can couple to the lattice topology in L-dependent ways. We report only the even-L results in the main text.

Seeds and uncertainty estimation. For each (L, T) point, observables are averaged over at least six independent seeds, equally split between the $g = +1$ and $g = -1$ flip sectors. Uncertainties on $\chi_{\max}(L)$ are estimated by bootstrap resampling over seeds. The effective exponent γ/ν and its uncertainty are extracted from a weighted least-squares fit of $\log \chi_{\max}$ versus $\log L$, with weights $w_i = 1/\sigma_i^2$, where $\sigma_i = \sigma_{\chi,i}/\chi_{\max,i}$ is the propagated 1σ uncertainty on $\log \chi_{\max}(L_i)$. The reported uncertainty on γ/ν is the standard error of the slope returned by this weighted fit.

Hardware reproducibility. All production runs were executed on identical hardware (NVIDIA RTX 5090 GPUs) to eliminate architecture-dependent variability in LLM inference. We observed during preliminary testing that runs of the same model with identical seeds and prompts can yield slightly different logits when executed on different GPU architectures, due to non-associativity of floating-point arithmetic and architecture-specific kernel implementations. While these differences

are negligible for most LLM applications, they can compound over the many micro-updates in our simulations and produce small but systematic shifts in sensitive quantities such as susceptibility and its peaks $\chi_{\max}$. Restricting all production runs to identical hardware ensures that finite-size scaling comparisons across different L values are not contaminated by this source of variability.

LLM inference calls. Across all three models, the full study comprised approximately 10^9 individual LLM inference calls. For each (L, T, seed) run, the number of calls is $N \times 2L^2 = 2L^4$ (one call per site per micro-update, with $N = L^2$ sites and $2L^2$ sweeps), so each lattice size contributes $240L^4$ calls when averaged over 6 seeds and scanned over 20 temperatures. This yields $\sim 1.4 \times 10^8$ calls for `llama3.1:8b` ($L \in \{8, 10, 12, 16, 20, 24\}$), $\sim 1.4 \times 10^8$ for `phi4-mini:3.8b` ($L \in \{10, 12, 16, 20, 24\}$), and $\sim 7.0 \times 10^8$ for `mistral:7b` ($L \in \{5, 10, 15, 20, 25, 30, 35\}$). This is the count for production runs only; exploratory runs bring the actual total somewhat higher. This computational cost is the primary reason for restricting finite-size scaling to roughly half a decade in L.