Calibrated but Autonomous: Inference-Time Bayesian Logit Correction for LLM Social Simulations

Ruggero Marino Lazzaroni*, **Lorenz Prattes***, **Jana Lasser**

IDea_Lab, University of Graz
Graz, Austria
{ruggero.lazzaroni, lorenz.prattes, jana.lasser}@uni-graz.at

Abstract

LLM-based social simulations, especially of social media, are a promising and growing area of research. How these simulations are designed varies widely across different axes, while issues regarding their empirical calibration persist. In this work, we categorize different approaches to action selection in Generative Agent-Based Models (GABMs) and put forward a novel approach that utilizes inference-time logit correction via Bayesian prior substitution. We test each approach via simulating a subreddit for a day and find our approach to produce more calibrated action distributions than the autonomous baselines, while at the same time seemingly preserving agent autonomy. We argue that this combination of calibration and context sensitivity, which none of the other approaches can achieve, is necessary for faithful and flexible social simulations, including counterfactual experimentation.

1 Introduction

In recent years, LLMs have been highly impactful in many fields of research. Among them, they have been theorized to have potential in research involving human behavior, such as fields that aim to simulate or predict behaviors of individuals or collectives (Bail 2023; Kozlowski and Evans 2025; Anthis et al. 2025; Ziems et al. 2024).

In this sphere, LLMs have been utilized as a new paradigm for Agent Based Model-like social simulations, also called Generative Agent Based Models (from now on, GABMs). Despite their high potential and growing number of examples in the literature (Park et al. 2023; Törnberg et al. 2023; Gao et al. 2023; Yang et al. 2024; Piao et al. 2025; Liu et al. 2025), recent studies have highlighted issues in validating such simulations in order to ensure their representativeness (Larooij and Törnberg 2026; Li and Tao 2026; Wang et al. 2025).

In particular, evidence suggests that, while empirical calibration is generally considered to be important for the validity of agent based simulations (Windrum, Fagiolo, and Moneta 2007; Larooij and Törnberg 2026), GABMs are systematically difficult to calibrate (Anthis et al. 2025; Li and Tao 2026).

One factor that has been largely unaddressed is that of action probabilities: which agents take which actions, and with what frequency. Different approaches exist: from discounting empirical calibration altogether and trusting LLM agents' behavioral distributions (Park et al. 2023; Yang et al. 2024; Gao et al. 2023; Qiao et al. 2025), to partially imposing choices via stochastic and statistically grounded distributions in some aspects of the simulation design, necessarily toning down the agency of the simulated users (Törnberg et al. 2023; Jeon et al. 2025).

While several of these systems externally control agent *activation* (i.e., which agents are active at a given timestep) through empirical distributions (Yang et al. 2024; Törnberg et al. 2023; Mou, Wei, and Huang 2024), *action selection* (i.e., what the active agent does) can be left to the LLM agent's free choice.

In this paper, we put forward an approach to solve this tension, by providing a method that enables informing LLM decision making by biasing them towards the empirical distributions while maintaining autonomous decision making by the agent. We then showcase this approach on a Social Media Simulation scenario to test its efficacy against strong baselines.

Specifically, our core contributions are as follows:

1. **A taxonomy of action selection in GABMs.** We formalize the distinction between puppeteered, uninformed autonomous, informed autonomous, and Bayesian approaches to action selection in social simulation.
2. **Testing the limits of explicit prompting.** We evaluate the "Informed Autonomous" condition, testing whether explicitly including numerical empirical rates in an agent's prompt can calibrate its action distribution for social media action selection. We demonstrate that, while this can partially mitigate the issue, it fails to reliably shape behavior, extending recent findings on LLMs' inability to follow probabilistic instructions (Gu et al. 2025; Misaki and Takase 2025) and on belief-behavior inconsistency in LLM role-playing (Mannekote et al. 2025).
3. **Logit-level prior correction.** We propose an inference-time Bayesian intervention that corrects action priors at the logit level, adapting the logit adjustment framework from long-tailed classification (Menon et al. 2021). A tunable strength parameter β interpolates between pure

*These authors contributed equally.

puppeteering ($\beta = 0$) and full context-sensitive correction ($\beta = 1$), successfully aligning agent behaviors with target empirical distributions without requiring fine-tuning or sacrificing the agents' autonomous decision-making.

2 Background

Recent research offers numerous examples of Generative Agent-Based Models applied to general social behavior (Park et al. 2023, 2024; Vezhnevets et al. 2023) and social media simulation specifically (Törnberg et al. 2023; Gao et al. 2023; Tang et al. 2024). By focusing on how these agents' action choices are (or are not) calibrated, we can distinguish four primary approaches.

Puppeteered Approach

Some simulations constrain agent behavior by externally sampling behavioral variables before invoking the LLM, which is then limited to generating content conditioned on these pre-set choices. For example, Jeon et al. (2025) use an ABM to pre-sample both the agent's stance and interaction type for each exchange, with the LLM only producing text that reflects these externally imposed conditions. Similarly, Park et al. (2022) programmatically generates users and interaction structures from a designer-specified blueprint.

Autonomous Approach

The most common approach in current literature is to grant agents near-complete behavioral autonomy. While agents are typically still activated based on a stochastic process and grounded in empirically derived or synthetic persona characterizations, they are free to select their actions based on their context, including their persona, memory, and environmental information fed to them. In the case of social media simulations, this usually corresponds to a social media feed: this is the approach employed, for example, by Yang et al. (2024), where agents choose from 21 available action functions with no biasing in the prompt. Other works adopting this paradigm include Park et al. (2023); Gao et al. (2023); Qiao et al. (2025); Liu et al. (2025); Piao et al. (2025).

Informed Autonomous Approach

Given the need for empirical grounding in social simulations, some studies have experimented with including explicit action rates in the agent prompt. Qiu et al. (2025) include retweet, quote, and rewrite rates in a zero-shot prompt condition and find that their exclusion leads to the largest drop in action prediction accuracy. Nonetheless, even with rates included, accuracy reaches only approximately 42%. More generally, studies have tested whether LLMs can self-calibrate based on probabilities verbalized in their context. Gu et al. (2025) prompt LLMs with explicit user activity levels and find that, while models can understand probabilities, they struggle with probability sampling at the precision required for behavioral simulation. Misaki and Takase (2025) formalize this limitation as a failure in "Probabilistic Instruction Following" (PIF), proposing a prompting fix that, while effective in isolation, requires complex multi-step reasoning per call, making it impractical for large-scale multi-agent simulation. Furthermore, Mannekote et al. (2025) show that LLMs exhibit systematic inconsistencies between their stated beliefs and their role-playing outputs, with individual-level forecasting accuracy degrading over longer horizons. Taubenfeld et al. (2024) demonstrate that RLHF-induced biases in action choices persist despite explicit instructions, and Lu et al. (2025) report that prompt-only methods achieve only 11.9% accuracy in generating human actions.

Calibration Approaches

Empirical calibration of LLM-based social simulations is a major concern in the field. Anthis et al. (2025) identify diversity in response distributions as a central challenge, finding that LLMs produce highly uniform outputs where humans exhibit wide variance. Li and Tao (2026) warn that plausible-looking simulation output does not imply faithfulness, with outcomes at risk of being "brittle, irreproducible, or overconfident." While specialized architectures can achieve accurate action distribution matching, as Zhang et al. (2025) report non-significant t-test deviations from ground-truth across views, likes, comments, and shares using group agents that each represent a collection of individuals, and Mi et al. (2025) reduce KL divergence to human distributions by 47% through mean-field fine-tuning, these approaches require either non-standard agent designs or model training. Huang et al. (2026) align agent behavior via a combined SFT and DPO pipeline, but calibrating individual user behavior remains a challenge even when data and compute are available, as training signals tend to reward predictive or preference-aligned outputs rather than faithful decision-making under realistic constraints (Li et al. 2024; Li and Tao 2026). The approach proposed in this work operates instead at inference time via logit manipulation, following the theoretical foundations of logit adjustment for long-tailed classification (Menon et al. 2021) and its extension to zero-shot foundation models (Zhu et al. 2023). It has parallels in Merchant and Levy (2025), who guide generation by adjusting the logits of a base model using auxiliary models, and in MF-LLM (Mi et al. 2025), which optimizes logits via gradient descent; however, this method concentrates probability mass to reduce uncertainty, whereas our approach redistributes it towards empirically grounded action rates. Therefore, our method occupies a middle ground: it mitigates the model's intrinsic action selection biases while anchoring the distribution to empirically calculated user-specific rates, ensuring the agent retains a degree of autonomous, context-sensitive influence over the final action chosen, all without fine-tuning or requiring multiple LLM invocations per turn.

3 Method

Bayesian Logit Correction

We propose a training-free, inference-time intervention that corrects the action selection distribution of LLM agents by operating directly on the model's output logits. The core idea

is to decompose the LLM's action prediction into a *context-sensitive* component, reflecting the model's semantic judgment about how a specific user would react to a specific feed, and a *prior* component, reflecting the model's baseline tendency for that user, shaped by pretraining and alignment biases rather than by actual behavior. We then discard the latter and replace it with the user's empirical action rates.

Let $a \in \mathcal{A}$ be an action (e.g., `post` or `comment`), c the content context (feed, thread), and u the user (persona, history). The LLM's implicit posterior over actions decomposes via Bayes' theorem into a content likelihood $\mathbb{P}_{\text{LLM}}(c \mid a, u)$, i.e., how well the content fits a user taking this action, and an action prior $\mathbb{P}_{\text{LLM}}(a \mid u)$, i.e., the model's bias towards picking certain actions given the user, which we believe models fail to calibrate properly. We construct a calibrated posterior by retaining the likelihood but replacing the prior with empirical rates $\pi_a = \mathbb{P}_{\text{data}}(a \mid u)$. In logit space, this yields:

$$\ell_{\text{cal}}(a) = \ell_{\text{LLM}}(a) - \log \mathbb{P}_{\text{LLM}}(a \mid u) + \log \pi_a \quad (1)$$

To control how much context-sensitive signal passes through, we introduce a strength parameter $\beta \in [0, 1]$:

$$\ell_{\text{final}}(a) = \beta \cdot \big(\ell_{\text{LLM}}(a) - \log \mathbb{P}_{\text{LLM}}(a \mid u)\big) + \log \pi_a \quad (2)$$

At β=0, the action is sampled directly from empirical rates (theoretically equivalent to a puppeteering approach). At β=1, the full Bayesian correction applies: the model can deviate from empirical rates based on context, but deviations are anchored to real behavior. The full formulation is detailed in Appendix A.

The model's implicit prior $\mathbb{P}_{\text{LLM}}(a \mid u)$ is estimated once per user before the simulation by prompting the model with randomly sampled feed snapshots and aggregating the resulting constrained action choices (see Appendix B). The correction is implemented as a logits processor within the vLLM inference engine (Kwon et al. 2023), intervening at the single token position where the model selects between action tokens.

4 Empirical Simulations

To evaluate differences in action probabilities between models and approaches, we construct a GABM-based simulation of a selected subreddit and its discussions. We instantiate four different experimental conditions that vary the degree and quality of autonomous action selection. Agents are contextualized with a part of the post and comment history of users active on the subreddit and tasked with continuing the discussions by creating posts or comments. Depending on the simulation scenario, they are presented with a respective subreddit feed or comment thread at the respective timestep. Additionally, agents can pick a reply target in comment threads, in order to form organic comment trees. At the start of the simulation, the feed and content still originate from the subreddit data, but as the simulation progresses, the percentage of generated content in the feed increases.

We make a number of simplifying assumptions. First, the action space is limited to posting and commenting, as these are the content-production actions for which per-user trace data is available; engagement actions such as upvoting are excluded as no individual-level data is observable. Second, we approximate the Reddit feed to 25 posts in reverse chronological order, as we lack data on per-user algorithmic ranking or upvote behavior. We acknowledge that this simplification will lessen some emergent dynamics such as virality, but it suffices to reveal structural differences in action selection behavior across conditions, which is our primary claim. Third, when prompt length exceeds a fixed token budget, we preferentially elide older feed items and earlier history entries to preserve recency as well as influential content. Finally, we limit the simulation to a highly active subset of users for whom we can reliably estimate empirical action rates from the training period.

Data and Model Selection

We select a subreddit for which a high percentage of users was relatively active, while fitting our size constraints on unique contributors ($5{,}000 < |U| < 500{,}000$) to allow for multiple simulations given our compute budget. We avoid subreddits whose topics heavily depend on evolving current events, and settle on `r/DebateReligion`. We retain the smallest set of users that account for at least 90% of actions taken, filtering out low-activity users for whom empirical action-rates cannot be reliably estimated.

We select a relatively recent time period of discussions to reduce the likelihood of its contents being included in training data for the models used, spanning January-June 2025, with a held-out evaluation window consisting of the 24 hours after the cutoff (i.e., July 1st, 2025). The training data is used to compute per-user activation rates and empirical action probabilities.

The models used in this evaluation are `Llama-3.1-8B-Instruct` (Grattafiori, Dubey et al. 2024) and `Qwen3.5-4B` (Qwen Team 2026), both open-source and selected to represent two models of major open-source model families, with a parameter count that enables affordable large scale social simulations. We limit this investigation to non-thinking versions.

Activation

We compute the daily action rate over the full training data, then distribute it across hours using the user's empirical circadian weight to create hourly activation probabilities. At each simulation tick, agents are activated via a Poisson process, which, given the current hour h and the agent's hourly rate λ_h, produces a per-tick activation probability. Our experiments run for a single simulated day divided into 288 ticks (5-minute bins).

Conditions

The experimental conditions contrast four different methods of agentic action selection, while the environments remain stable. At each activation, an agent must select an action $a \in \mathcal{A} = \{\texttt{post}, \texttt{comment}\}$. If commenting is selected, a second turn presents the relevant thread. The conditions differ only in how a is determined.

Puppeteered. The puppeteered condition does not involve the LLM in the selection step and directly instructs the agent on the action it will take, sampled from the user's empirical action distribution π. This serves as an upper bound that preserves distribution accuracy but removes any agency from the decision-making. This also precludes counterfactual experimentation, as the sampled rates remain fixed regardless of changes to the simulated environment.

Autonomous. The LLM agent selects a without any external biasing, based solely on its persona and the presented feed in its context. This in principle aligns with the approach followed in most existing GABMs, such as Yang et al. (2024).

Informed-Autonomous. Similarly to the autonomous approach, agents are free to choose, but are informed of π by including it in the prompt as natural language (e.g., "post: 15.0%, comment: 85.0%"). This tests whether informing the model of behavioral statistics is sufficient for it to reproduce them.

Bayesian. The agent selects a as in the autonomous condition, but its output logits are corrected at the point of selection using the Bayesian prior correction described in §3. The empirical rates π do not appear in the prompt; instead, the correction operates directly on logits, biasing them according to priors estimated in a pre-simulation step (Appendix B).

Simulation Design

Each condition is simulated for one day across three random seeds for each of the two models, varying the stochastic activation sequence. We compare the four action selection conditions described above: puppeteered, autonomous, informed-autonomous, and our Bayesian approach. For the Bayesian condition, we additionally evaluate multiple values of the strength parameter $\beta \in \{0.25, 0.5, 1.0\}$.

5 Results and Discussion

Action Rate Accuracy

As a preliminary evaluation, we test how accurately each condition can reproduce a given target post rate in isolation. We select 5 real users from the subreddit with varying activity levels and assign each a set of synthetic target post rates spanning $\{1\%, 5\%, 10\%, 25\%, 50\%\}$, covering the realistic range of user behaviors. For each user-rate pair, we generate 50 action decisions (250 total per condition), each prompted with the user's real profile and a randomly sampled feed. We compare the autonomous, informed-autonomous, and Bayesian conditions, with $\beta \in \{0.0, 0.25, 0.5, 1.0\}$ for the latter. Accuracy is measured as mean absolute error (MAE) between observed and target post rates across all 25 user-rate combinations (Table 1).

Both prompt-based conditions fail to track target rates. The autonomous condition reflects each model's fixed action bias: Qwen defaults to $\sim$10-15% posting and Llama to $\sim$22-26% regardless of user or target. The informed-autonomous condition shows only marginal improvement, confirming that LLMs cannot reliably self-calibrate from verbalized probabilities.

Condition	Qwen3.5-4B	Llama-3.1-8B
Autonomous	0.145	0.180
Informed-autonomous	0.132	0.116
Bayesian β=0.0	**0.029**	0.048
Bayesian β=0.25	0.072	**0.021**
Bayesian β=0.5	0.075	0.041
Bayesian β=1.0	0.111	0.058

Table 1: Mean absolute error between observed and target post rates across 5 users $\times$ 5 target rates. Lower is better. Bold indicates best per model.

The Bayesian correction substantially outperforms both baselines. At β=0 the correction achieves near-perfect rate matching, meeting the expectations; as β increases, the model's contextual preferences are progressively readmitted, trading rate fidelity for context sensitivity. The optimal β differs by model: Qwen's stronger intrinsic bias requires lower β for accurate calibration, while Llama's more uniform priors tolerate higher values. Full per-target-rate breakdowns are reported in Appendix C. Wilcoxon signed-rank tests confirm that the Bayesian condition significantly outperforms the informed-autonomous baseline across most β values ($p < 0.01$), with the exception of β=1.0 on Qwen ($p = 0.18$), where the full context-sensitive correction approaches the error of the informed prompt-based baseline.

Simulation Results

Action Rate Analysis. Considering simulation results in the calibration of action distributions, we evaluate each condition's ability to reproduce the empirical distribution of actions, reporting both Jensen-Shannon divergence (JSD) between simulated and ground-truth per-user action distributions and per-user mean absolute error (MAE) on post rates, averaged across seeds (Table 2).

Table 2: Simulation Action distribution accuracy by condition and model. JSD and MAE reported as mean $\pm$ std across seeds. Lower is better. Full details available in Appendix E

	Qwen 3.5-4B		Llama 3.1-8B	
Condition	JSD	MAE	JSD	MAE
Autonomous	.187±.021	.271±.041	.124±.008	.162±.022
Inf.-auton.	.149±.072	.198±.148	.063±.005	.040±.006
Puppeteered	.064±.009	.021±.004	.067±.002	.022±.005
Bay. β=0.25	.068±.004	.021±.001	.067±.001	.022±.005
Bay. β=0.5	.066±.007	.017±.003	.066±.006	.023±.003
Bay. β=1.0	**.063**±.007	**.016**±.002	.066±.003	.027±.003

The autonomous condition is the worst performer on both metrics for both models, confirming that uncalibrated LLM agents substantially deviate from empirical action distributions. Including explicit rates in the prompt (informed-

autonomous) partially mitigates the issue for Llama but remains unstable for Qwen (JSD 0.149 $\pm$ 0.072), and in neither case approaches the accuracy of the mechanically grounded approaches.

The Bayesian correction matches puppeteered accuracy across all β values for both models ($\sim$0.065 JSD, $\sim$0.02 MAE). Notably, for Qwen, β=1.0, which retains full context sensitivity, achieves the best JSD and MAE of any condition including puppeteered, suggesting that when the prior is well-estimated, the model's contextual judgment can actively improve calibration rather than degrading it.

Emergent Dynamics. Furthermore, we examine the effects of different conditions on structural dynamics observable in the simulation. We focus on two emergent properties that are not directly targeted by any condition's design: temporal clustering of posts and comment thread depth.

Temporal Burstiness. The temporal burstiness is measured by the Fano factor (ratio of variance to mean) of binned posting counts over time, shown in Figure 1.

The autonomous conditions exhibit pronounced burstiness (Fano $\approx$ 4-13). The puppeteered condition, by contrast, produces near-uniform temporal distributions (Fano $<$ 1), consistent with its reliance on independent Poisson activation with no content-driven decision-making. The Bayesian conditions occupy a middle ground, with Fano factors close to 1 across both models (range: 0.85-1.09). This is consistent with the design intent to anchor the overall rate, by suppressing activation patterns observed in autonomous conditions. The context-sensitive component ($\beta > 0$) allows a mild temporal variation, although in practice the variation remains small, potentially due to a lack of beta calibration.

Comment thread depth. We additionally analyze the depth distribution of comment threads, where depth 1 denotes a top-level reply to a post and higher values indicate nesting of reply chains, as shown in Figure 2. In the ground truth, threads reach a mean depth of 6.3 with a long tail extending beyond depth 80, and 47% of comments occur at depth 4 or greater.

All simulated conditions produce substantially more shallow threads (mean depth 1.0-1.9, max depth 4-14), with the majority of comments remaining at depth 1. This is in part attributable to the feed dynamics, where engagement is not accounted for. However, the conditions that allocate more actions to commenting, Bayesian and puppeteered, consistently produce deeper threads than the autonomous baselines, as a higher volume of comments creates more opportunities for nested replies.

Agency Preservation. While the Bayesian correction achieves distributional accuracy comparable to puppeteering (Table 2), it differs in a key aspect: puppeteered actions are sampled independently of feed content, whereas the Bayesian correction at $\beta > 0$ lets the model's contextual judgment influence action selection. We illustrate this with two qualitative examples from the simulation runs.

User A (empirical post rate: 1.65%) produced zero posts across all activations under the puppeteered regimen in both models. Under the Bayesian condition, they posted four times across all β values and seeds and all four posts concerned Islam, consistent with the user's profile and posting history. Both Llama and Qwen independently generated Islam-related posts for this user when the feed contained relevant discussion topics, suggesting that the action choice was driven by contextual relevance rather than random sampling.

User B (empirical post rate: 0.23%) similarly produced zero puppeteered posts. Under the Bayesian condition, each model produced one post: on the role of symbolism in religious belief (Llama) and on the ontological status of the soul (Qwen). While the topics differed, both were philosophically substantive and consistent with the user's history of engagement with abstract theology.

These examples, although not systematic, show the advantage of the Bayesian approach over puppeteering: actions are taken *when contextually appropriate*. A puppeteered simulation matches the empirical distribution by design but cannot capture the content-dependent nature of human action selection: the reason a real user posts is not a coin flip, but a response to what they see in their feed. This property is also essential for counterfactual simulation designs, where researchers may alter the environment and need agents to adjust their behavior in response while maintaining calibrated base rates: a puppeteered agent would act at the same rate regardless of the intervention, while a Bayesian agent can react to the changed conditions while remaining anchored to empirical priors.

6 Limitations and Future Work

We recognize that the simulations presented in this work serve as a proof of concept for the proposed approach rather than a comprehensive evaluation. The scope is limited temporally (a single simulated day), by number of users (a filtered subset of one subreddit), and by model variety (two open-source LLMs at comparable parameter scales). We reserve for future work the execution of longer simulations spanning multiple days or weeks, with larger and more diverse user populations across different subreddits and the calibration of beta values per model.

Furthermore, we acknowledge that the agency preservation analysis (Section 5) is not systematic and does not quantify content-action alignment across the full user population. Developing metrics for contextual appropriateness remains necessary to substantiate this method beyond a proof of concept.

It should be considered, additionally, that the action space in our evaluation is limited to posting and commenting, as these are the content-production affordances for which per-user trace data is available on Reddit. A natural extension would be to incorporate engagement actions such as upvoting and downvoting, which would require either platform-level aggregate data or alternative data sources that expose individual voting behavior. More broadly, the method generalizes to any scenario in which an LLM agent must select among discrete actions at empirically grounded rates: like social media simulations with richer action spaces (e.g., sharing, reacting, reporting), dialogue simulations where

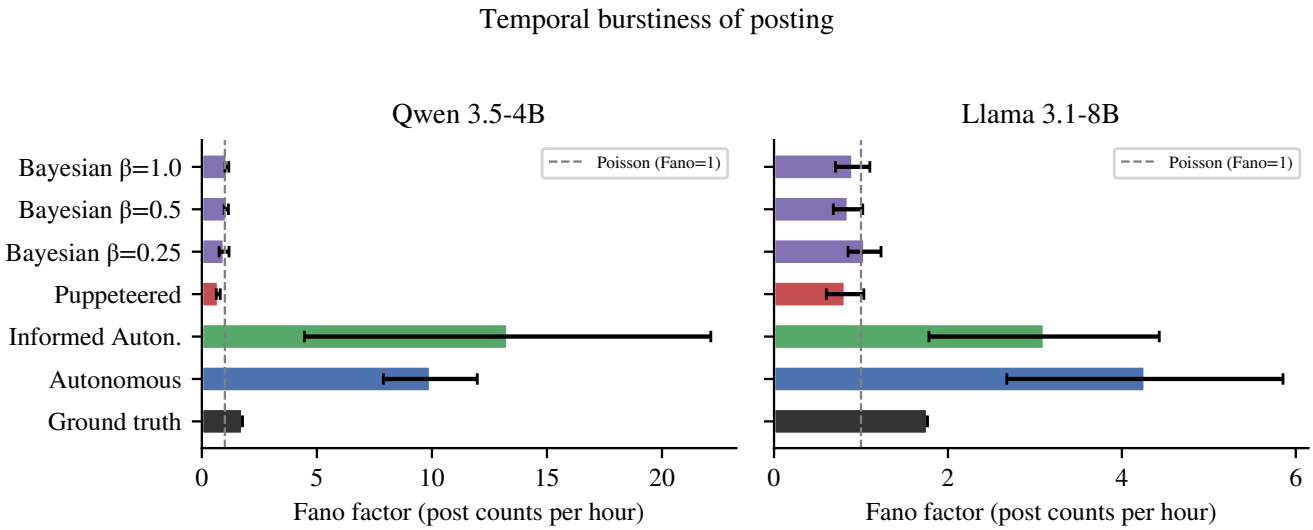

Figure 1: **Temporal burstiness of posting activity across conditions and models.** We report the Fano factor (ratio of variance to mean) of post counts in one hour bins. A Fano factor of 1 indicates equal likelihood of an event over the bins, while higher factors indicate bursts.

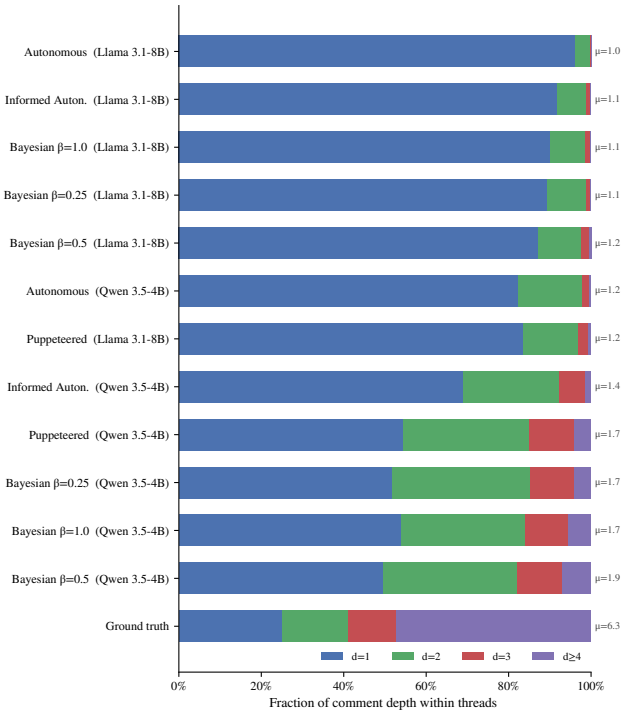

Figure 2: **The comment thread depth variation across conditions and models.** The colored bars depict the different depths of comments emerging in the simulation. While the difference to the ground truth is large for all models, the Bayesian variants generally show more depth compared to their autonomous counterparts.

agents must choose between response strategies or multi-agent economic simulations where agents select among trading actions.

Beyond action selection in GABMs, the Bayesian logit correction can in principle be applied to any task in which an LLM is considered semantically capable of understanding the context but requires or can benefit from calibration of its output distribution against an external prior. Potential applications include survey simulation, where LLM respondents must reproduce known demographic response distributions; recommendation systems, where a model's content preferences need to be re-anchored to observed user engagement patterns; or clinical decision support, where a model's diagnostic suggestions could be informed by population-level base rates rather than its training distribution. In each case, the core assumption holds: the model's *relative* preferences across options are informative, but its *absolute* rates are miscalibrated and can be corrected at inference time without retraining. The ability to substitute arbitrary priors for empirical ones also makes the method naturally suited for counterfactual experimentation, where researchers need agents to enact hypothetical behavioral shifts while preserving context-sensitive decision-making.

7 Conclusion

In this work, we introduced a taxonomy of action selection strategies in Generative Agent-Based Models and proposed Bayesian logit correction, an inference-time intervention that calibrates LLM agent action distributions toward empirical rates without fine-tuning or sacrificing autonomous decision-making. Across two open-source models and four experimental conditions on a Reddit social simulation, the correction matched the distributional accuracy of puppeteered agents while preserving the model's ability

to condition action choices on context. Our results further demonstrate that explicitly informing agents of target rates via prompting is insufficient to reliably shape their behavior, reinforcing recent findings on the limits of probabilistic instruction following in LLMs. While the scope of our evaluation is constrained, the method could be further generalized to any setting in which an LLM must select among discrete actions according to empirically grounded rates. We consider its extension to richer action spaces, longer time horizons, and systematic evaluation of agency preservation a promising direction for future work in this field.

Acknowledgments

Ruggero Marino Lazzaroni and Jana Lasser have received funding from the European Research Council (ERC) under the European Union's Horizon Europe programme (Grant agreement No. 101160928).

References

Anthis, J. R.; Liu, R.; Richardson, S. M.; Kozlowski, A. C.; Koch, B.; Brynjolfsson, E.; Evans, J.; and Bernstein, M. S. 2025. Position: LLM Social Simulations Are a Promising Research Method. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*.

Bail, C. A. 2023. Can Generative AI Improve Social Science? *Science*, 380(6650): 1108–1109.

Gao, C.; Lan, X.; Lu, Z.; Mao, J.; Piao, J.; Wang, H.; Jin, D.; and Li, Y. 2023. S^3: Social-network Simulation System with Large Language Model-Empowered Agents. *arXiv preprint arXiv:2307.14984*.

Grattafiori, A.; Dubey, A.; et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Gu, J.; Pang, L.; Shen, H.; and Cheng, X. 2025. Do LLMs Play Dice? Exploring Probability Distribution Sampling in Large Language Models for Behavioral Simulation. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, 5375–5390. Abu Dhabi, UAE: Association for Computational Linguistics.

Huang, R.; Tang, N.; Xu, J.; Cao, Y.; Tu, Q.; Guo, S.; Zheng, B.; Liu, H.; and Yang, Y. 2026. PolicySim: An LLM-Based Agent Social Simulation Sandbox for Proactive Policy Optimization. arXiv:2603.19649.

Jeon, M. S.; Mendoza, M.; Fernández, M.; Providel, E.; Rodríguez, F.; Espina, N.; Carvallo, A.; and Abeliuk, A. 2025. Simulating Conversations on Social Media with Generative Agent-Based Models. *EPJ Data Science*, 14(79).

Kozlowski, A. C.; and Evans, J. 2025. Simulating Subjects: The Promise and Peril of Artificial Intelligence Stand-Ins for Social Agents and Interactions. *Sociological Methods & Research*.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J. E.; Zhang, H.; and Stoica, I. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Larooij, M.; and Törnberg, P. 2026. Validation is the Central Challenge for Generative Social Simulation: A Critical Review of LLMs in Agent-Based Modeling. *Artificial Intelligence Review*, 59(15).

Li, K.; Dai, C.; Zhou, W.; and Hu, S. 2024. Fine-Grained Behavior Simulation with Role-Playing Large Language Model on Social Media. *arXiv preprint arXiv:2412.03148*.

Li, Y.; and Tao, D. 2026. Position: AI Agents Are Not (Yet) a Panacea for Social Simulation. *arXiv preprint arXiv:2603.00113*.

Liu, G.; Rahman, S.; Kreiss, E.; Ghassemi, M.; and Gabriel, S. 2025. MOSAIC: Modeling Social AI for Content Dissemination and Regulation in Multi-Agent Simulations. *arXiv preprint arXiv:2504.07830*.

Lu, Y.; Huang, J.; Han, Y.; Yao, B.; Bei, S.; Gesi, J.; Xie, Y.; Zheshen; Wang; He, Q.; and Wang, D. 2025. Can LLM Agents Simulate Multi-Turn Human Behavior? Evidence from Real Online Customer Behavior Data. arXiv:2503.20749.

Mannekote, A.; Davies, A.; Li, G.; Boyer, K. E.; Zhai, C.; Dorr, B. J.; and Pinto, F. 2025. Do Role-Playing Agents Practice What They Preach? Belief-Behavior Consistency in LLM-Based Simulations of Human Trust. *arXiv preprint arXiv:2507.02197*.

Menon, A. K.; Jayasumana, S.; Rawat, A. S.; Jain, H.; Veit, A.; and Kumar, S. 2021. Long-tail Learning via Logit Adjustment. In *International Conference on Learning Representations (ICLR)*.

Merchant, H.; and Levy, B. 2025. A Fast and Effective Solution to the Problem of Look-ahead Bias in LLMs. *arXiv preprint arXiv:2512.06607*.

Mi, Q.; Yang, M.; Yu, X.; Zhao, Z.; Deng, C.; An, B.; Zhang, H.; Chen, X.; and Wang, J. 2025. MF-LLM: Simulating Population Decision Dynamics via a Mean-Field Large Language Model Framework. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Misaki, K.; and Takase, S. 2025. String Seed of Thought: Prompting LLMs for Distribution-Faithful and Diverse Generation. *arXiv preprint arXiv:2510.21150*.

Mou, X.; Wei, Z.; and Huang, X. 2024. Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 4789–4809. Bangkok, Thailand: Association for Computational Linguistics.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM.

Park, J. S.; Popowski, L.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology (UIST)*. ACM.

Park, J. S.; Zou, C. Q.; Shaw, A.; Hill, B. M.; Cai, C.; Morris, M. R.; Willer, R.; Liang, P.; and Bernstein, M. S. 2024. Generative Agent Simulations of 1,000 People. *arXiv preprint arXiv:2411.10109*.

Piao, J.; Yan, Y.; Zhang, J.; Li, N.; Yan, J.; Lan, X.; Lu, Z.; Zheng, Z.; et al. 2025. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. *arXiv preprint arXiv:2502.08691*.

Qiao, B.; Li, K.; Zhou, W.; Li, S.; Lu, Q.; and Hu, S. 2025. BotSim: LLM-Powered Malicious Social Botnet Simulation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 39, 14377–14385.

Qiu, Z.; Lyu, H.; Xiong, W.; and Luo, J. 2025. Can LLMs Simulate Social Media Engagement? A Study on Action-Guided Response Generation. *arXiv preprint arXiv:2502.12073*.

Qwen Team. 2026. Qwen3.5: Towards Native Multimodal Agents.

Tang, J.; Gao, H.; Pan, X.; Wang, L.; Tan, H.; Gao, D.; Chen, Y.; Chen, X.; Lin, Y.; Li, Y.; Ding, B.; Zhou, J.; Wang, J.; and Wen, J.-R. 2024. GenSim: A General Social Simulation Platform with Large Language Model based Agents. *arXiv preprint arXiv:2410.04360*.

Taubenfeld, A.; Dover, Y.; Reichart, R.; and Goldstein, A. 2024. Systematic Biases in LLM Simulations of Debates. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 251–267. Miami, Florida, USA: Association for Computational Linguistics.

Törnberg, P.; Valeeva, D.; Uitermark, J.; and Bail, C. 2023. Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. *arXiv preprint arXiv:2310.05984*.

Vezhnevets, A. S.; Agapiou, J. P.; Aharon, A.; Ziv, R.; Matyas, J.; Duéñez-Guzmán, E. A.; Cunningham, W. A.; Osindero, S.; et al. 2023. Generative Agent-Based Modeling with Actions Grounded in Physical, Social, or Digital Space Using Concordia. *arXiv preprint arXiv:2312.03664*.

Wang, Q.; Wu, J.; Tang, Z.; Luo, B.; Chen, N.; Chen, W.; and He, B. 2025. What Limits LLM-based Human Simulation: LLMs or Our Design? *arXiv preprint arXiv:2501.08579*.

Windrum, P.; Fagiolo, G.; and Moneta, A. 2007. Empirical Validation of Agent-Based Models: Alternatives and Prospects. *Journal of Artificial Societies and Social Simulation*, 10(2): 8.

Yang, Z.; Zhang, Z.; Zheng, Z.; Jiang, Y.; Gan, Z.; Wang, Z.; Ling, Z.; Chen, J.; et al. 2024. OASIS: Open Agent Social Interaction Simulations with One Million Agents. *arXiv preprint arXiv:2411.11581*.

Zhang, Y.; Song, Z.; Zhou, H.; Ren, W.; Chen, Y.-P. P.; Yu, J.; and Yang, W. 2025. GA-S^3: Comprehensive Social Network Simulation with Group Agents. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Zhu, B.; Tang, K.; Sun, Q.; and Zhang, H. 2023. Generalized Logit Adjustment: Calibrating Fine-tuned Models by Removing Label Bias in Foundation Models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Ziems, C.; Held, W.; Shaikh, O.; Chen, J.; Zhang, Z.; and Yang, D. 2024. Can Large Language Models Transform Computational Social Science? *Computational Linguistics*, 50(1): 237–291.

A Bayesian Formulation

Let $a \in \mathcal{A}$ be an action, c a token sequence representing the content context, and u a token sequence representing the user. Applying Bayes' theorem with u as background conditioning:

$$\mathbb{P}(a \mid c, u) \propto \mathbb{P}(c \mid a, u) \cdot \mathbb{P}(a \mid u) \tag{3}$$

The LLM produces its own estimate of this posterior, which decomposes analogously into a content likelihood $\mathbb{P}_{\text{LLM}}(c \mid a, u)$ representing the model's semantic judgment of how well the context matches the user taking this action, and an action prior $\mathbb{P}_{\text{LLM}}(a \mid u)$ capturing the model's baseline expectation, shaped by pretraining and alignment rather than by actual user behavior.

We isolate the likelihood by dividing out the implicit prior, then substitute the user's empirical action rate $\pi_a = \mathbb{P}_{\text{data}}(a \mid u)$:

$$\mathbb{P}_{\text{cal}}(a \mid c, u) \propto \frac{\mathbb{P}_{\text{LLM}}(a \mid c, u)}{\mathbb{P}_{\text{LLM}}(a \mid u)} \cdot \pi_a \tag{4}$$

Since $\mathbb{P}_{\text{LLM}}(a \mid c, u) \propto \exp(\ell_a)$, this becomes additive in logit space:

$$\ell_{\text{cal}}(a) = \ell_{\text{LLM}}(a) - \log \mathbb{P}_{\text{LLM}}(a \mid u) + \log \pi_a \tag{5}$$

We introduce $\beta \in [0, 1]$ to scale the context-dependent term. This yields the final formula (formula 2 in the main text):

$$\ell_{\text{final}}(a) = \beta \cdot \big(\ell_{\text{LLM}}(a) - \log \mathbb{P}_{\text{LLM}}(a \mid u)\big) + \log \pi_a \tag{6}$$

Note that $\log \pi_a$ is not scaled by $(1-\beta)$: the empirical anchor is always present, and β controls only the magnitude of context-dependent deviations from it. At β=0 the formula reduces to $\log \pi_a$ (puppeteering); at β=1 the full correction applies.

Approximation with reasoning. When the model generates a reasoning trace r before the action token, logits become $\ell_{\text{LLM}}(a \mid c, u, r)$ and the clean prior decomposition is approximate since r may correlate with a. The correction remains well-motivated as logit adjustment (Menon et al. 2021), in which shifting logits by $\log \pi_a - \log \mathbb{P}_{\text{LLM}}(a \mid u)$ corrects the model's marginal action distribution toward empirical rates regardless of the conditioning context. The key assumption, that the model's *relative* action preferences are informative but its *absolute* rates are miscalibrated, holds with or without reasoning.

B Prior Estimation

The model's implicit prior $\mathbb{P}_{\text{LLM}}(a \mid u)$ is estimated per-user before the simulation:

1. Sample $N{=}10$ feed snapshots uniformly from the user's training period.
2. For each snapshot, prompt the model with the user's persona, the sampled feed and instructions about the task, constraining generation (after a trigger token is detected for parsing purposes) to the action token set $\mathcal{A}$ by masking all other tokens to $-\infty$. No correction is applied, so the formula reduces to a constrained binary choice on raw logits.
3. Aggregate choices with Laplace smoothing: $\hat{\mathbb{P}}_{\text{LLM}}(a \mid u) = (\text{count}(a) + 1) \,/\, (N + |\mathcal{A}|)$.

Constrained generation is used rather than logprob extraction from unconstrained generation, as in the full vocabulary, the bare action tokens carry negligible probability mass (<0.01%), due to the model distributing probability across space-prefixed, capitalized, and other surface variants. Normalizing two near-zero values yields meaningless ratios. Since softmax over two logits depends only on their difference, constrained generation forces a binary choice using the model's full contextual understanding.

C Full Results of Action Rate Accuracy Test

Tables 3 and 4 report the observed post rate for each target rate and condition. Each cell aggregates 250 action decisions (50 per user $\times$ 5 users).

Target	Auton.	Inf.-aut.	β=1	β=.5	β=.25	β=0
1%	9.6	5.2	6.8	5.6	3.6	**0.8**
5%	14.8	5.2	11.6	10.0	6.8	**5.2**
10%	11.2	3.6	16.0	13.2	12.4	**6.8**
25%	9.6	6.8	15.3	21.6	21.2	**22.4**
50%	16.4	15.6	23.6	32.4	31.2	**52.4**

Table 3: Observed post rate (%) by condition for Qwen3.5-4B. Bold indicates closest to target.

Target	Auton.	Inf.-aut.	β=1	β=.5	β=.25	β=0
1%	22.8	2.4	0.4	0.8	**0.8**	0.8
5%	26.0	2.8	5.6	4.0	5.6	**5.2**
10%	24.0	2.8	12.8	10.4	**9.6**	8.8
25%	25.2	1.2	32.4	29.6	**24.8**	24.8
50%	21.2	28.0	48.0	51.6	**49.6**	46.4

Table 4: Observed post rate (%) by condition for Llama-3.1-8B. Bold indicates closest to target.

D Used Prompts

Action Selection Prompt (Autonomous / Bayesian Condition)

Note: <activity_stats> is only displayed in the Informed-Autonomous Condition.

```
SYSTEM:
You are simulating a Reddit user.

<activity_stats>
  post: 23.5%
  comment: 76.5%
</activity_stats>

Decide what this user does next. Output ONLY valid XML, nothing else.

If posting:
<action>post</action>
<generated_post>
<title>...</title>
<body>...</body>
</generated_post>

If commenting on a feed post:
<action>comment</action>
<feed>F3</feed>

<user_history>
u/catholiccrusader77

[P1] r/debatereligion · "Is moral relativism defensible?"
[P2] r/debatereligion · "The problem of evil revisited"
[P3] r/debatereligion · "Why do Muslims believe the hadiths are valid?"
  ...
</user_history>

<feed>
[F1] r/debatereligion · u/bob · 5 comments · "Is moral relativism defensible?"
[F2] r/debatereligion · u/charlie · 2 comments · "The problem of evil revisited"
[F3] r/debatereligion · u/alice · 0 comments · "Free will and determinism"
  ...
[+12 posts]
                                        <- elision placeholder
</feed>

USER:
What does this user do next?
```

Comment Reply Prompt (Turn 2, Optional, All Conditions)

```
SYSTEM:
You are simulating a Reddit user.

Write a comment this user would plausibly add to the thread.
Output ONLY valid XML, nothing else.

<target>[root] or C-number</target>
<reply>...</reply>
```

```
<user_history>
u/catholiccrusader77
[P1] r/debatereligion · "Is moral relativism defensible?"
  ...
</user_history>

<thread>
[root] r/debatereligion · u/bob · [title] "Is moral relativism defensible?"
"In this thread we discuss whether there's objective morality..."
  [C1] u/charlie · "I'd argue most people are relativists..."
    [C2] u/diana · "But then how do we judge harmful practices?"
      [C3] u/evan · "That's the key question!"
  [C4] u/frank · "Actually, I think objectivity is..."
</thread>

USER:
Write your comment now.
```

Post Generation Prompt (Puppeteered Condition)

```
    SYSTEM:
You are simulating a Reddit user.

Generate a submission this user would plausibly post next.
Output ONLY valid XML, nothing else.

<generated_post>
<title>...</title>
<body>...</body>
</generated_post>

<user_history>
u/alice
[P1] r/debatereligion · "What are the strongest arguments for utilitarianism?"
  ...
</user_history>

<feed>
[F1] r/debatereligion · u/bob · 5 comments · "Is moral relativism defensible?"
  ...
</feed>

USER:
Write your post now.
```

E Simulation Action Distribution Details

Table 5: Posting probability (fraction of actions that are posts) by condition and model. Simulated values reported as mean ± std across seeds.

Condition	Qwen 3.5-4B	Llama 3.1-8B
Autonomous	.194±.155	.030±.006
Inf.-auton.	.268±.040	.150±.013
Puppeteered	.009±.003	.009±.002
Bay. β=0.25	.007±.001	.009±.003
Bay. β=0.5	.005±.002	.011±.003
Bay. β=1.0	.004±.001	.015±.001
Ground truth (core)	.007	
Training (core)	.009	