# Spiral of Silence in Large Language Model Agents

**Mingze Zhong[1], Meng Fang[2], Zijing Shi[1], Yuxuan Huang[2], Shunfeng Zheng[1], Yali Du[3], Ling Chen[1], Jun Wang[4]**
[1]AAII, University of Technology Sydney, NSW, Australia
{Mingze.Zhong, Zijing.Shi, Shunfeng.Zheng}@student.uts.edu.au
Ling.Chen@uts.edu.au
[2] University of Liverpool, Liverpool, UK
{Meng.Fang, Yuxuan.Huang}@liverpool.ac.uk
[3] King's College London, London, UK
yali.du@kcl.ac.uk
[4] University College London, London, UK
jun.wang@ucl.ac.uk

## Abstract

The Spiral of Silence (SoS) theory holds that individuals with minority views often refrain from speaking out for fear of social isolation, enabling majority positions to dominate public discourse. When the "agents" are large language models (LLMs), however, the classical psychological explanation is not directly applicable, since SoS was developed for human societies. This raises a central question: can SoS-like dynamics nevertheless emerge from purely statistical language generation in LLM collectives? We propose an evaluation framework for examining SoS in LLM agents. Specifically, we consider four controlled conditions that systematically vary the availability of "History" and "Persona" signals. Opinion dynamics are assessed using trend tests such as Mann–Kendall and Spearman's rank, along with concentration measures including kurtosis and interquartile range. Experiments across open-source and closed-source models show that history and persona together produce strong majority dominance and replicate SoS patterns; history signals alone induce strong anchoring; and persona signals alone foster diverse but uncorrelated opinions, indicating that without historical anchoring, SoS dynamics cannot emerge. The work bridges computational sociology and responsible AI design, highlighting the need to monitor and mitigate emergent conformity in LLM-agent systems.

## 1 Introduction

In human societies, public opinion is shaped by complex social dynamics. The *Spiral of Silence* (SoS) theory (Noelle-Neumann, 1974) posits that individuals tend to withhold opinions they perceive as unpopular due to fear of social isolation. This creates a self-reinforcing cycle in which minority views gradually disappear from public discourse, making dominant opinions appear increasingly widespread.

A pressing question is whether analogous dynamics can arise in populations of artificial agents. Large language models (LLMs) are rapidly deployed in multi-agent environments where they collaborate, negotiate, or compete (OpenAI, 2023; Touvron et al., 2023). Unlike humans, LLMs do not experience emotions or social anxieties, yet recent studies show they adapt to social-like cues—displaying sycophantic tendencies toward users (Sharma et al., 2023) and reproducing group-level biases (Ashery et al., 2024). If SoS-like effects can emerge in such settings, this would challenge traditional emotion-based explanations and raise critical concerns about bias amplification, conformity, and opinion manipulation in LLM-driven systems.

To investigate this possibility, we design an evaluation framework for detecting SoS dynamics in LLM agents. We adopt a controlled movie-rating task, which provides a quantifiable environment for measuring opinion formation. Two key signals are introduced: (i) *History*, the average rating of preceding agents, serving as a dynamic proxy for the collective opinion climate; (ii) *Persona*, a role assigned to each agent that encodes predispositions and ensures diversity of initial preferences. Crossing these signals allows us to disentangle the influence of collective anchoring and individual predisposition, and to test whether their interaction gives rise to SoS-like convergence.

We hypothesize that SoS dynamics will be most evident when they drive agents to align with perceived majorities. To measure this, we combine concentration statistics such as interquartile



*Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23238–23253

range (Clark-Carter, 2005) and kurtosis (Balanda and MacGillivray, 1988) with trend diagnostics including the Mann–Kendall test (Mann, 1945) and Spearman's rank correlation (Spearman, 1904). We evaluate a range of LLM families, including the open-source `Qwen` (Bai et al., 2023), `DeepSeek` (Liu et al., 2024), and `Mistral`, as well as the closed-source `GPT-4o-mini` (Hurst et al., 2024), enabling both cross-family comparisons and within-family scaling analyses. Our findings show that in the absence of social signals, agents default to positive movie ratings; persona signals promote opinion heterogeneity; history signals exert strong anchoring effects; and SoS dynamics emerge most clearly when both signals are present.

Our contributions are as follows:

- We propose a systematic framework for testing SoS in LLM agents, isolating the roles of collective influence through History and individual predisposition through Persona.[1]
- We provide the first evidence that SoS-like dynamics can arise from learned language generation mechanisms, yielding new insights into the nature of conformity.
- We identify behavioral regularities such as positivity bias and anchoring effects, and discuss their implications for the design and governance of multi-agent AI systems.

## 2 Methodology

This section describes the task simulation, presents the evaluation framework, and specifies the metrics and criteria for measuring opinion dynamics and detecting SoS effects in LLM agents.

### 2.1 Problem Setup

We construct a multi-agent environment to simulate an online rating system. In this task, a population of $N$ agent sequentially rate the same movie on an integer scale. We formalize the rating space as an $M$-level cardinal metric: $\mathcal{M} \triangleq \{1, \ldots, M\}$. At step $k$, the rating given by agent $i$ to movie $j$ is denoted by $r_{j,k} \in \mathcal{M}$. The set of historical ratings for movie $j$ up to the $k$-th rating is defined as $\mathcal{H}_{j,k} \triangleq \{r_{j,1}, \ldots, r_{j,k}\}$.

As agents rate sequentially, each observes prior ratings giving its own. The "collective opinion climate", representing an agent's perception of public opinion, is defined at step $k+1$ as the average of all preceding ratings available to the $(k+1)$-th agent: $\mathcal{F}(\mathcal{H}_{j,k}) = \frac{1}{k}\sum_{l=1}^{k} r_{j,l}$.

### 2.2 LLM Agent Design

We design LLM agents by varying the presence of *History* and *Persona* signals to examine their effects on the emergence of SoS.

- **History:** We operationalize the collective opinion climate as the average rating of all preceding agents, provided as input to the next agent. This signal is endogenous and dynamic: each new rating updates the climate and is passed forward, creating a feedback loop. Such recursive updating allows a slight majority to amplify its influence under the SoS effect, distinguishing this process from a static anchoring effect.
- **Persona:** To introduce heterogeneous predispositions, each agent is assigned a unique persona. Personas are specified through rich textual descriptions covering attributes such as occupation, interests, and background.

This $2 \times 2$ design yields four controlled scenarios:

- **History + Persona**: The agent is assigned a persona and observes the historical average rating of preceding agents. This condition captures how an identity-driven agent behaves under the influence of a collective opinion climate.
- **History only**: The agent observes only the historical average rating, without a persona. This isolates the effect of perceived public opinion on a generic agent.
- **Persona only**: The agent receives only a persona description, providing a fixed identity and predispositions but no historical signal. This condition examines how internal preferences, shaped by persona, influence rating behavior.
- **No History, No Persona**: The agent receives only rating instructions and movie information, without persona context or historical ratings. This baseline captures a generic agent's behavior in the absence of external signals.

The corresponding prompt templates are provided in Appendix A.1, and an example persona is given in Appendix A.2.

### 2.3 Quantifying SoS Dynamics

To evaluate whether LLM agents exhibit Spiral of Silence behavior, we focus on two defining patterns: (i) a dynamic self-reinforcing process whereby the majority opinion becomes increasingly dominant

[1]Data and code available at: https://github.com/aialt/SoS-LLMs

over time, and (ii) a final outcome of high consensus, where minority opinions have effectively faded. To capture these aspects, we employ opinion trend metrics and rating concentration metrics.

#### 2.3.1 Opinion Trend Metrics

We introduce the Majority-Conforming Opinion (MCO) sequence, which captures the dominant opinion trend by aggregating cumulative proportions of positive and negative ratings. Formally, for movie $j$ at $k$-th rating, we compute the cumulative proportions of positive and negative ratings as $\text{pos}_{j,k} = \frac{1}{k}\sum_{t\le k}\mathbf{1}\big[r_{i,j,t} \ge 6\big]$ and $\text{neg}_{j,k} = \frac{1}{k}\sum_{t\le k}\mathbf{1}\big[r_{i,j,t} \le 5\big]$, where $\mathbf{1}[\cdot]$ is the indicator function. The MCO sequence is then defined as $\text{MCO}_{j,k} = \max\big\{\text{pos}_{j,k}, \text{neg}_{j,k}\big\}$. On this basis, two metrics are proposed to quantify opinion trends.

- **Mann–Kendall statistic** ($S$), which is well-suited for detecting monotonic trends in time series without assuming linearity. We apply the Mann–Kendall trend test to the MCO sequence. The $S$ statistic is defined as $S_j = \sum_{k=m}^{T-1}\sum_{t=k+1}^{T}\text{sgn}\big(\text{MCO}_{j,t} - \text{MCO}_{j,k}\big)$, where $m$ denotes the starting round for analysis and $T$ is the final round of ratings. A significantly positive $S$ shows that the majority opinion strengthens monotonically over time. This provides direct evidence for the self-reinforcing dynamic that characterizes the SoS.
- **Spearman's Rank Correlation** ($\rho$), which is used to quantify the strength of a monotonic trend. To complement this, we compute Spearman's rank correlation between the MCO sequence and the time steps. It is calculated as $\rho_j = 1 - \frac{6\sum d_k^2}{n(n^2-1)}$, where $d_k$ is the difference between the rank of the $k$-th time step and the rank of its corresponding $\text{MCO}_{j,k}$ value, and $n$ is the number of observations (i.e., $n = T-m+1$). A value of $\rho$ close to 1 indicates that the majority opinion exhibits a strong monotonic increase.

#### 2.3.2 Rating Concentration Metrics

To assess whether the ratings of LLM agents converge toward a strong consensus, we evaluate the dispersion of the final set of $L$ ratings using the following metrics:

- **Kurtosis**, which measures the "peakedness" of the rating distribution and indicates whether the ratings are sharply concentrated around a single value. We compute it over the final $L$ ratings as $\text{Kurt}_L(j) = \frac{1}{L}\sum_{k=T-L+1}^{T}\left(\frac{r_{i,j,k}-\mu_j}{\sigma_j}\right)^4 - 3$, where $\mu_j$ and $\sigma_j$ are the mean and standard deviation of the same $L$ ratings. A positive Kurtosis value indicates a distribution more sharply peaked than the normal distribution, showing that ratings are concentrated around a single value. This pattern aligns with the consensus formation expected under the SoS.
- **Interquartile Range (IQR)**, which provide a robust measure of the central spread of recent ratings. We use it to assess whether late-stage ratings are tightly clustered, with a smaller IQR indicating stronger opinion concentration. It is calculated as $IQR_L(j) = Q_3^{(L)}(j) - Q_1^{(L)}(j)$, where $Q_1^{(L)}(j)$ and $Q_3^{(L)}(j)$ denote the 25th and 75th percentiles of the last $L$ ratings.

## 3 Experiments

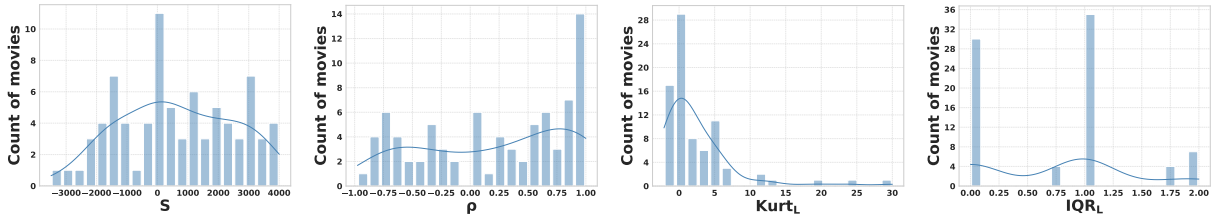
### 3.1 Experimental Setup

**Models.** We evaluate a range of backend LLMs, including one closed-source model, GPT-4o-mini, and six open-source models: DeepSeek-V2-Lite-Chat, Mistral-8B-Instruct, and Qwen-2.5 series (1.5B, 3B, 7B).

**Dataset.** We construct our dataset from two main sources to support the evaluation framework.

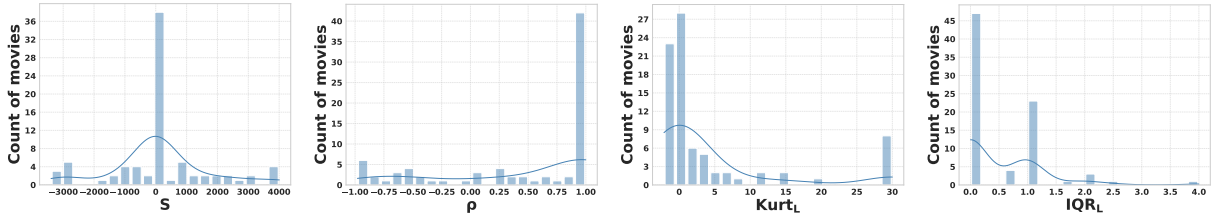
- **Movies:** We scrape movie data from IMDb covering films released after January 12, 2025, including titles, genres, overviews, and IMDb average scores. This cutoff date is selected to ensure that the movies fall outside the models' training data, thereby preventing data contamination.
- **Personas:** We randomly sample 100 distinct profiles from the `elite_persona` subset of the PersonaHub dataset (Ge et al., 2024). This subset is chosen for its coherent and information-rich descriptions, which support persona-based agent modeling.

**Implementation Details.** In our experiments, we use a rating scale of $M = 10$ and fix the agent population size at $N = 100$. For each movie, ratings are collected sequentially in a randomized order of agents. In the "w/ History" scenario, the prompt for the $n$-th agent includes the average rating of the preceding $n-1$ agents. The generation temperature is fixed at 0.1 in all experiments, and each agent's final rating is the average of three independent model runs. For evaluation, ratings $\ge 6$ are considered positive and ratings $\le 5$ negative when




(a) Mann–Kendall Distribution (b) Spearman Distribution (c) Kurtosis Distribution (d) IQR Distribution

Figure 1: Results of SoS metrics for GPT-4o-mini in Scenario I (History + Persona).




(a) Mann–Kendall Distribution (b) Spearman Distribution (c) Kurtosis Distribution (d) IQR Distribution

Figure 2: Results of SoS metrics for Mistral-8b-instruct in Scenario I (History + Persona).

computing option trend metrics. For rating concentration metrics, we use $L = 30$ final ratings, and we prepend $m = 10$ randomly generated warm-up ratings to each sequence to mitigate early-stage volatility.

### 3.2 Results

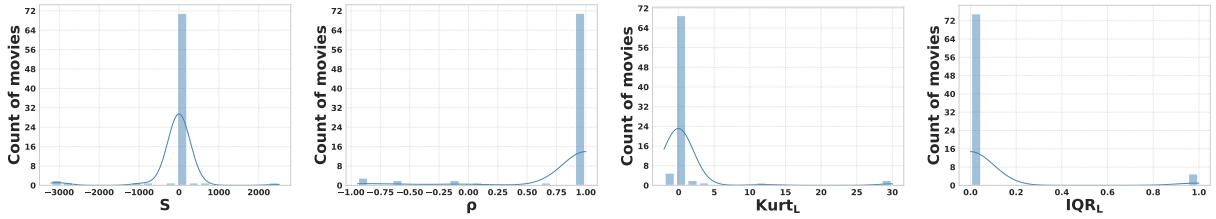
This section present detailed results for two representative models, the closed-source GPT-4o-mini and the open-source Mistral-8B-Instruct, across all four experimental scnarios. For the complete set of results covering all other models, please refer to Appendix A.3.

**Scenario I (History + Persona): SoS emerges when both history and persona signals are present.** The metric distributions for GPT-4o-mini in Figure 1 and for Mistral-8B-Instruct in Figure 2 show a clear progression from early disagreement to final consensus when both history and persona signals are present. This pattern provides strong evidence for the SoS effect, reflected in two key dimensions:

- **Opinion Trend:** Both models exhibit a positive monotonic trend. As shown in Figure 1a, GPT-4o-mini's Mann-Kendall ($S$) statistic tends to be positively skewed, and its Spearman's rank correlation ($\rho$) has a strong peak near 1.0 as shown in Figure 1b. In Figures 2a and 2b, Mistral-8B-Instruct shows an even more pronounced trend, with its ($\rho$) distribution concentrated at 1.0. These results indicate that majority opinions undergo a gradually reinforcing and monotonic convergence process, consistent with the SoS effect.
- **Rating Concentration:** Both models' ratings indicate a high degree of final consensus. As shown in Figures 1c and 1d, GPT-4o-mini exhibits a positively skewed kurtosis distribution and an IQR distribution with distinct peaks near 0 and 1. Mistral-8B-Instruct displays a similar pattern, as illustrated in Figures 2c and 2d, with kurtosis skewed toward higher values and IQR values heavily concentrated at low values. These distributions suggest a substantial reduction in opinion diversity at the final stage.

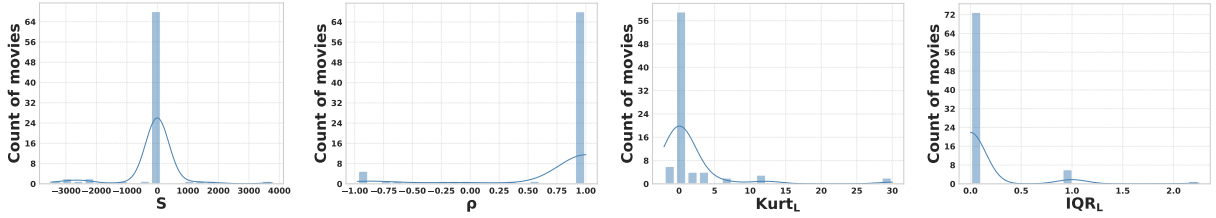
**Scenario II (History only): Anchoring collective opinion appears when only history signals are present.** When only the history signal is available , the rating distributions for GPT-4o-mini and Mistral-8B-Instruct, as shown in Figures 3 and 4, exhibit a clear anchoring effect. This can be observed from two aspects.

- **Opinion Trend:** The lack of dynamic opinion evolution is evident in both models. The Mann-Kendall ($S$) distributions, as shown in Figures 3a and 4a, show an extremely sharp peak around 0, indicating a lack of any monotonic trend. The Spearman values ($\rho$) are heavily concentrated around 1.0, as shown in Figures 3b and 4b. These patterns reflects a strong anchoring effect, where initial opinions constrain later ratings, leading to minimal variation throughout the sequence.




(a) Mann–Kendall Distribution (b) Spearman Distribution (c) Kurtosis Distribution (d) IQR Distribution

Figure 3: Results of SoS metrics for GPT-4o-mini in Scenario II (History Only).




(a) Mann–Kendall Distribution (b) Spearman Distribution (c) Kurtosis Distribution (d) IQR Distribution

Figure 4: Results of SoS metrics for Mistral-8b-instruct in Scenario II (History Only).

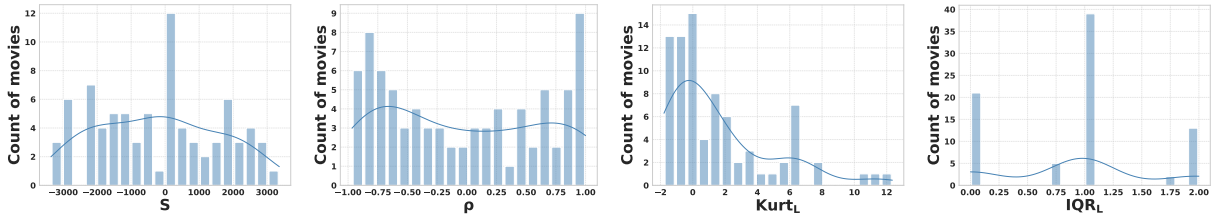
- **Rating Concentration:** The concentration metrics clearly illustrate the outcome of anchoring effect. As shown in Figures 3c and 3d for GPT-4o-mini, and Figures 4c and 4d for Mistral-8B-Instruct, both models converge to a state of extreme and stable consensus. The IQR values are heavily concentrated at 0, indicating minimal variation across agents, while the strongly positive kurtosis values reflect peaked, tightly clustered rating distributions. These patterns confirm that once early opinions are established, later responses remain fixed, resulting in highly uniform outcomes.

**Scenario III (Persona only): Promoting opinion diversity appears when only persona signals are present.** With only persona signals , the distributions for GPT-4o-mini and Mistral-8B-Instruct, as shown in Figure 5 and in Figure 6, show that both models foster opinion diversity without exhibiting any clear convergence.

- **Opinion Trend:** As shown in Figures 5a and 5b, and Figures 6a and 6b, the distributions of the Mann–Kendall ($S$) and Spearman's rank correlation ($\rho$) statistics for both models show no clear trend. The values are broadly spread and centered around 0, indicating that opinions vary across agents and do not follow a consistent directional pattern.
- **Rating Concentration:** Figures 5c and 5d, and 6c and 6d show that the concentration metrics support the maintenance of opinion diversity. For both models, the Kurtosis distribution is centered near 0 and includes negative values, while the IQR distributions are dispersed across a range of values, indicating that opinions remain heterogeneous.

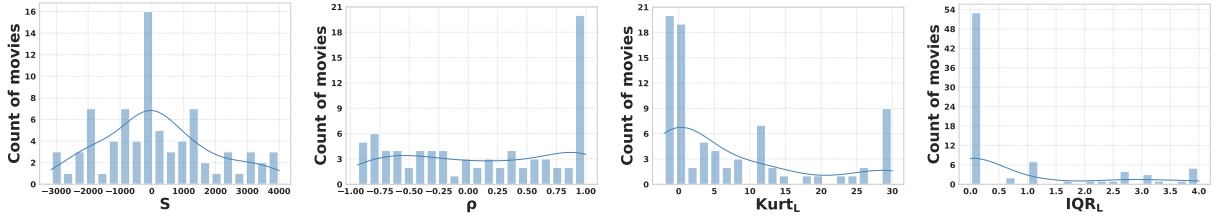
**Scenario IV (No History, No Persona): Revealing inherent biases under baseline, namely, when history and persona signals are not present.** This scenario excludes both persona and history signals , serving as a baseline to capture the models' intrinsic tendencies. In this setting, the distributions for GPT-4o-mini and Mistral-8B-Instruct, as shown in Figure 7 and Figure 8, reveal an inherent bias.

- **Trend Metrics:** Both models display static and highly uniform trends. As shown in Figures 7a and 7b, and Figures 8a and 8b, the Spearman's rank correlation ($\rho$) is sharply peaked at 1.0, while the Mann–Kendall ($S$) values are narrowly centered around 0. This suggests that the model defaults to a consistent opinion regardless of the movie, reflecting an internal bias rather than a context-driven response.
- **Concentration Metrics:** As shown in Figures 7c and 7d, and Figures 8c and 8d, both models display extreme rating concentration. The interquartile range (IQR) values are sharply peaked at 0, and the kurtosis values are strongly positive, indicating a highly uniform and narrowly distributed set of ratings across agents.




(a) Mann–Kendall Distribution (b) Spearman Distribution (c) Kurtosis Distribution (d) IQR Distribution

Figure 5: Results of SoS metrics for GPT-4o-mini in Scenario III (Persona Only).




(a) Mann–Kendall Distribution (b) Spearman Distribution (c) Kurtosis Distribution (d) IQR Distribution

Figure 6: Results of SoS metrics for Mistral-8b-instruct in Scenario III (Persona Only).

### 3.3 Case Study

This section visualizes the evolution of positive and negative opinion proportions over time for a single movie under each experimental condition. We focus on GPT-4o-mini, which serves as a representative example consistent with the patterns observed in other models.

**Scenario I.** A typical case in the history + persona scenario is shown in Figure 9. Initially, opinions are diverse due to the different personas. However, once one side gains a slight advantage in the early rounds, its dominance rapidly strengthens, while the minority opinion is quickly suppressed and ultimately silenced. This perfectly replicates the self-reinforcing dynamic central to SoS theory.

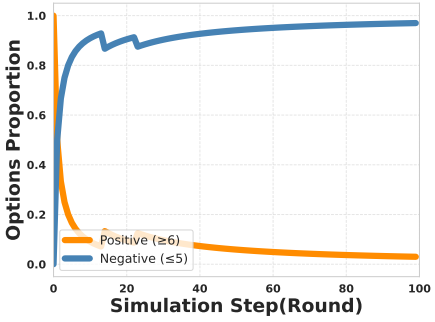
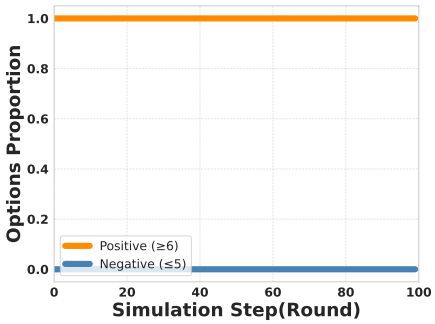
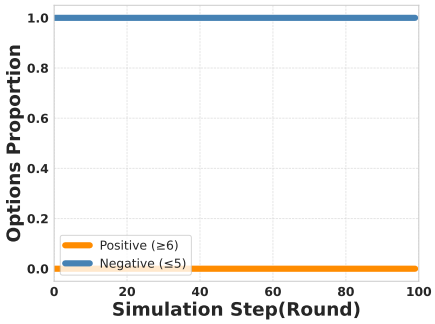



Figure 9: A case study of the movie "Norma" with GPT-4o-mini in Scenario I (History + Persona).

**Scenario II.** When only history signal is present, we observe a strong anchoring effect rather than a self-reinforcing spiral. As shown in Figure 10a and Figure 10b, the collective opinion is randomly dominated by either positive or negative opinions at the start and then remains almost completely stable for the entire process. The opinion proportion remains unchanged over time, showing no signs of dynamic evolution.

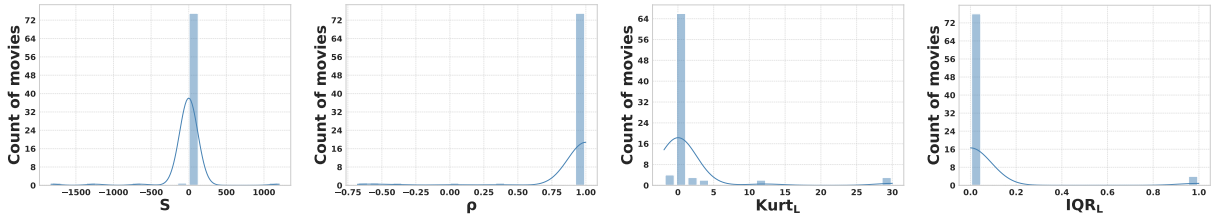



(a) A Positive Case.



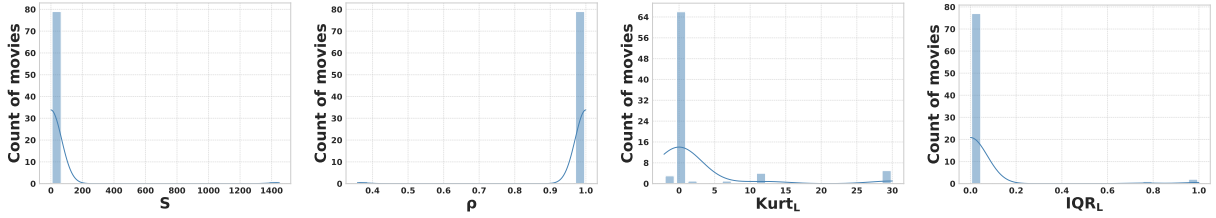

(b) A Negative Case.

Figure 10: A positive case study of the movie "Azzad" and a negative case study of the movie "Norma" with GPT-4o-mini in Scenario II (History Only).




(a) Mann–Kendall Distribution (b) Spearman Distribution (c) Kurtosis Distribution (d) IQR Distribution

Figure 7: Results of SoS metrics for GPT-4o-mini in Scenario IV (No History, No Persona).




(a) Mann–Kendall Distribution (b) Spearman Distribution (c) Kurtosis Distribution (d) IQR Distribution

Figure 8: Results of SoS metrics for Mistral-8b-instruct in Scenario IV (No History, No Persona).

**Scenario III.** When only persona signal is present , the opinion distribution exhibits a distinct dynamic. As shown in Figure 11 positive and negative opinions fluctuate and compete throughout the rating process, with neither side achieving a lasting advantage.

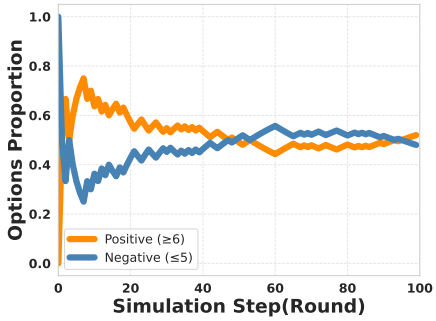



Figure 11: A case study of the movie "Norma" with GPT-4o-mini in Scenario III (Persona Only).

**Scenario IV.** In this scenario, when neither history nor persona signals are present, the model reveals its inherent bias. As shown in Figure 12, the proportion of positive opinions remains fixed at 1.0 from the very beginning, with negative opinions entirely absent throughout the process. This static pattern is not the result of opinion dynamics but rather reflects a built-in positivity prior, reflecting the model's default tendency to favor positive ratings when no contextual signals are available. This positivity bias aligns with prior LLM-as-a-judge evidence showing leniency toward higher scores and sycophancy effects that inflate positive ratings.

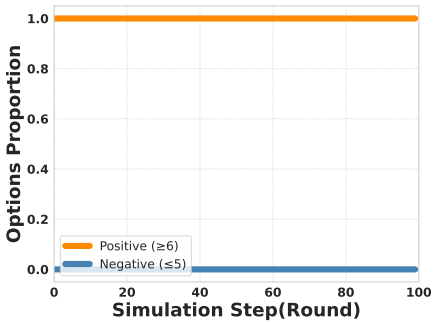



Figure 12: A case study of the movie "Azzad" with GPT-4o-mini in Scenario IV (No History, No Persona).

### 3.4 Persona-Context Consistency

We further explore whether an agent is more likely to conform to the collective opinion when its assigned persona aligns closely with the movie's content. To quantify this alignment, we introduce semantic match score and rating distance. The semantic match score quantifies the similarity between an agent's persona description and the movie's overview using TF-IDF–based cosine similarity. The rating distance is defined as the absolute difference between an agent's individual rating and the historical average of all prior ratings, capturing the degree of deviation from collective opinion:

$$\mathrm{Dist}(r_{i,j,k}) = \left| r_{i,j,k} - \mathcal{F}(\mathcal{H}_{j,k}) \right|. \quad (1)$$

The results, as shown in Figure 13, reveal a clear

negative correlation between semantic match score and rating distance. When the alignment between an agent's persona and the movie overview is low, ratings are more widely dispersed and frequently deviate from the collective opinion. In contrast, high semantic similarity corresponds to tightly clustered ratings with minimal variance. These findings suggest that consistency between persona and movie context plays a crucial role in shaping conformity. Agents whose personas are more closely aligned with the content are more likely to follow collective opinion. This pattern likely reflects greater confidence in its own assessment, resulting in more stable and less variable ratings.

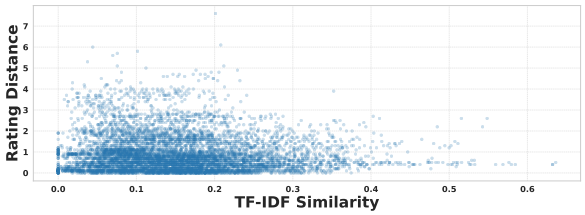



Figure 13: The relation between **Semantic Match vs. Rating Distance**.

## 4 Related Work

The SoS theory, introduced by Noelle-Neumann (1974), posits that individuals suppress minority opinions due to fear of social isolation. Subsequent work has tested these ideas online. For example, a Pew survey (Hampton et al., 2014) shows that social media users are much more likely to voice opinions when they believe that their network agrees with them, and Porten-Cheé and Eilders (2015) shows that anonymity and low-effort feedback significantly increase willingness to express unpopular views on online forums. More recently, researchers have explored how LLM agents can simulate such social dynamics (Chuang et al., 2023). Park et al. (2023) use generative agents in a simulated town; these agents exhibit emergent social behaviors. Similarly, Nasim et al. (2025) presents Gensim, a general social simulation platform with LLM agents, and Light et al. (2023); Shi et al. (2023) studies a community of LLMs playing the social deduction game Avalon. Akata et al. (2025) use behavioral game theory to let LLM agents play finitely repeated games, finding that models develop consistent cooperative or defection strategies. Sarkadi et al. (2019) introduces the Traitors framework for LLMs to study trust and deceit. Leng and Yuan (2023) analyzes LLM responses in canonical economics games using a probabilistic SUVA framework and reports that the decisions of most models reflect social welfare and reciprocity considerations rather than pure self-interest.

Other recent work focuses on how LLMs represent the majority opinions (Weng et al., 2025). For example, Ye et al. (2024) systematically quantifies biases in LLM-as-a-Judge and identifies a strong bandwagon effect. To explore the dynamics of opinion, Nasim et al. (2025) proposes a simulator that embeds LLM-based agents into networked opinion spread models. By integrating classic theories of social influence (Kelman, 1958; Munroe, 2013) with LLM communication, their framework allows researchers to study how LLM agents propagate influence. Likewise, (Yang et al., 2024) presents OASIS, an open-scale social media simulator with up to a million LLM agents, and shows that larger simulated populations produce richer group dynamics and greater opinion diversity, and Zhao et al. (2024) shows the diversity of LLM agents. These LLM-based platforms connect directly to prior work on collective behavior: classical agent-based models by (Deffuant et al., 2002; Rainer and Krause, 2002; Friedkin and Johnsen, 2011) demonstrate how repeated local interactions can produce global consensus or polarization.

## 5 Conclusion

Our study shows that LLM-based movie rating agents exhibit a clear positivity bias by default, yet develop more diverse opinions when given distinct personas and increasingly conform to prior context when a historical collective opinion is provided. By crossing binary signals design (persona × history), we isolate the influence of each signal: persona alone induces opinion diversity, history alone imposes anchoring consistency, and only their combination triggers a pronounced SoS. These results highlight that a SoS can spontaneously emerge in LLM agents without any emotional drive: purely from the interplay between internalized statistical biases and externally presented collective signals. This insight underscores the power of social context in shaping AI behavior and reminds us to remain alert to the social biases embedded in LLMs that can influence such simulations.

## 6 Limitations and Potential Risks

**Limitations.** Our study is subject to several practical constraints. First, due to available computational resources, our experiments focus on lightweight and midsized open-source models, rather than very large-scale models, which may exhibit different emergent dynamics. Second, our simulation of social feedback adopts a simplified agent, that is, providing agents only with the historical average rating as a substitute for social influence. Although this abstraction enables controlled investigation of majority dynamics, it does not capture the full range of factors shaping opinion formation in real-world societies, such as emotion, network structure, or identity effects. However, we believe that these design choices allow us to isolate and systematically analyze the core mechanisms of the emergence of the SoS in collectives of LLM-based agents, and we leave more complex extensions for future work.

**Potential Risks.** Our study shows that purely algorithmic LLM agents can reproduce SoS effect. Although this advances scientific understanding, it also entails several risks: Malicious actors could adapt our protocol to build large-scale manipulative campaigns or persuasive LLM-based chatbots that systematically nudge users toward the perceived majority, thus suppressing dissenting voices; If the initial prompt or training data carry demographic, political, or cultural biases, the SoS mechanism may magnify those biases and further marginalize minority opinions.

**Licenses.** All models and tools used in this study are released under open-source or research licenses.

## 7 Acknowledgements

We thank the anonymous reviewers for their insightful comments and constructive feedback.

## References

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2025. Playing repeated games with large language models. *Nature Human Behaviour*, pages 1–11.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2024. The dynamics of social conventions in llm populations: Spontaneous emergence, collective biases and tipping points. *arXiv preprint arXiv:2410.08948*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Kevin P Balanda and HL MacGillivray. 1988. Kurtosis: a critical review. *The American Statistician*, 42(2):111–119.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. 2023. Simulating opinion dynamics with networks of llm-based agents. *arXiv preprint arXiv:2311.09618*.

David Clark-Carter. 2005. Interquartile range. *Encyclopedia of Statistics in Behavioral Science*.

Guillaume Deffuant, Frédéric Amblard, Gérard Weisbuch, and Thierry Faure. 2002. How can extremism prevail? a study based on the relative agreement interaction model. *Journal of artificial societies and social simulation*, 5(4).

Noah E Friedkin and Eugene C Johnsen. 2011. *Social influence network theory: A sociological examination of small group dynamics*, volume 33. Cambridge University Press.

Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

Keith N Hampton, Harrison Rainie, Weixu Lu, Maria Dwyer, Inyoung Shin, and Kristen Purcell. 2014. Social media and the'spiral of silence'.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Herbert C Kelman. 1958. Compliance, identification, and internalization three processes of attitude change. *Journal of conflict resolution*, 2(1):51–60.

Yan Leng and Yuan Yuan. 2023. Do llm agents exhibit social behavior? *arXiv preprint arXiv:2312.15198*.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. 2023. Avalonbench: Evaluating llms playing the game of avalon. *arXiv preprint arXiv:2310.05036*.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, et al. 2024. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Henry B. Mann. 1945. Nonparametric tests against trend. *Econometrica*, 13(3):245–259.

Paul T Munroe. 2013. Social influence network theory: a sociological examination of small group dynamics.

Mehwish Nasim, Syed Muslim Gilani, Amin Qasmi, and Usman Naseem. 2025. Simulating influence dynamics with llm agents. *arXiv preprint arXiv:2503.08709*.

Elisabeth Noelle-Neumann. 1974. The spiral of silence a theory of public opinion. *Journal of communication*, 24(2):43–51.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.

Pablo Porten-Cheé and Christiane Eilders. 2015. Spiral of silence online: How online communication affects opinion climate perception and opinion expression regarding the climate change debate. *Studies in communication sciences*, 15(1):143–150.

Hegselmann Rainer and Ulrich Krause. 2002. Opinion dynamics and bounded confidence: models, analysis and simulation.

Ştefan Sarkadi, Alison R Panisson, Rafael H Bordini, Peter McBurney, Simon Parsons, and Martin Chapman. 2019. Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4):287–302.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Zijing Shi, Meng Fang, Shunfeng Zheng, Shilong Deng, Ling Chen, and Yali Du. 2023. Cooperation on the fly: Exploring language agents for ad hoc teamwork in the avalon game. *arXiv preprint arXiv:2312.17515*.

Charles Spearman. 1904. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. 2025. Do as we do, not as you think: the conformity of large language models. *arXiv preprint arXiv:2501.13381*.

Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, Bowen Dong, et al. 2024. Oasis: Open agents social interaction simulations on one million agents. *arXiv preprint arXiv:2411.11581*.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, et al. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.

Xiutian Zhao, Ke Wang, and Wei Peng. 2024. An electoral approach to diversify llm-based multi-agent collective decision-making. *arXiv preprint arXiv:2410.15168*.

# A Appendix

## A.1 LLM Agent Rating Prompts

We present the four distinct prompts employed in our experiments. Each prompt corresponds to one of the four experimental settings, differing by the presence or absence of historical average rating information, referred to as *History*, and character profile information, referred to as *Persona*. These prompts were designed to isolate and evaluate the specific contributions of social influence and persona-based individual differences to the emergence of the SoS effect.

**Prompt: History + Persona**

```
Please provide your rating for the movie.

# Your Character Profile
You are [persona]

# Movie Information
Title: [Movie Title]
Genres: [Genres]
Overview: [Movie Overview]
Movie average rating: [Historical Average] (1-10)

# Rating Principle
Rate the above movie on an integer scale from 1 to 10, where:

- 1 = Awful/Abysmal (unwatchable)
- 5 = Mediocre/Unsure (forgettable)
- 10 = Perfect/Masterpiece (flawless)

# Output Principle
Provide only a single integer (1-10) without extra text.
```

**Prompt: History only**

```
Please provide your rating for the movie.

# Movie Information
Title: [Movie Title]
Genres: [Genres]
Overview: [Movie Overview]
Movie average rating: [Historical Average] (1-10)

# Rating Principle
Rate the above movie on an integer scale from 1 to 10, where:

- 1 = Awful/Abysmal (unwatchable)
- 5 = Mediocre/Unsure (forgettable)
- 10 = Perfect/Masterpiece (flawless)

# Output Principle
Provide only a single integer (1-10) without extra text.
```

Prompt: Persona only

```
Please provide your rating for the movie.

# Your Character Profile
You are [persona]

# Movie Information
Title: [Movie Title]
Genres: [Genres]
Overview: [Movie Overview]

# Rating Principle
Rate the above movie on an integer scale from 1 to 10, where:

- 1 = Awful/Abysmal (unwatchable)
- 5 = Mediocre/Unsure (forgettable)
- 10 = Perfect/Masterpiece (flawless)

# Output Principle
Provide only a single integer (1-10) without extra text.
```

Prompt: No History, No Persona

```
Please provide your rating for the movie.

# Movie Information
Title: [Movie Title]
Genres: [Genres]
Overview: [Movie Overview]


# Rating Principle
Rate the above movie on an integer scale from 1 to 10, where:

- 1 = Awful/Abysmal (unwatchable)
- 5 = Mediocre/Unsure (forgettable)
- 10 = Perfect/Masterpiece (flawless)

# Output Principle
Provide only a single integer (1-10) without extra text.
```

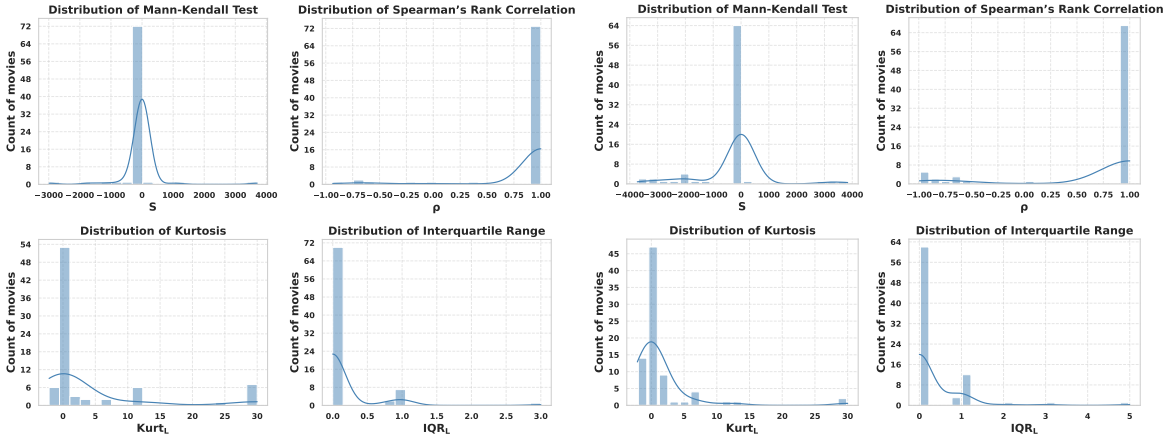
### A.2 Persona Example

```
A computer enthusiast who is interested in optimizing the performance of their system,
particularly the CPU, GPU, and RAM. They are looking for software tools that can help them
monitor and control the performance of their system, and they are willing to invest time
in learning how to use these tools effectively. They are not necessarily looking for a
professional-grade software tool, but rather a user-friendly and easy-to-use software that
can provide comprehensive information about their system's performance and help them optimize
it. They are also interested in software tools that can help them monitor the stability of
their system after overclocking, as they want to avoid damaging their system.
```

Figure 14: A persona example

### A.3 Metrics Distributions for Other Models

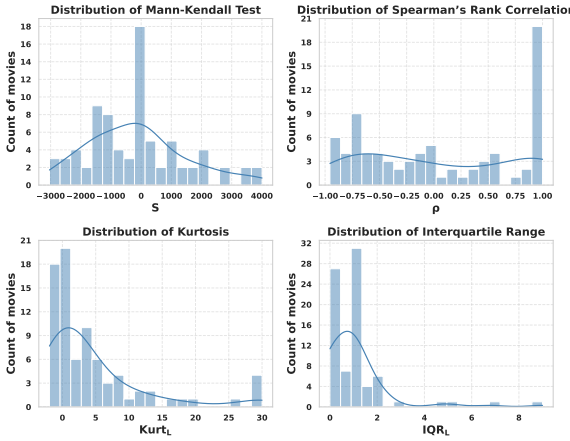
In this section, we present the distribution of statistical metrics for additional models including DeepSeek V2 Lite Chat and the Qwen 2.5 series (1.5B, 3B, and 7B) under each of the four experimental scenarios. For each model, we visualize and summarize four key statistics across all movies: the Mann–Kendall statistic ($S$), Spearman's rank correlation coefficient ($\rho$), late-stage interquartile range (IQR), and late-stage kurtosis.



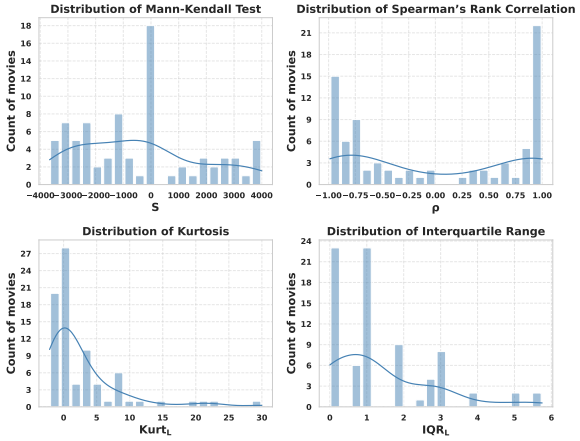

(a) "w/o Persona & w/o History"




(b) "w/o Persona & w/ History"

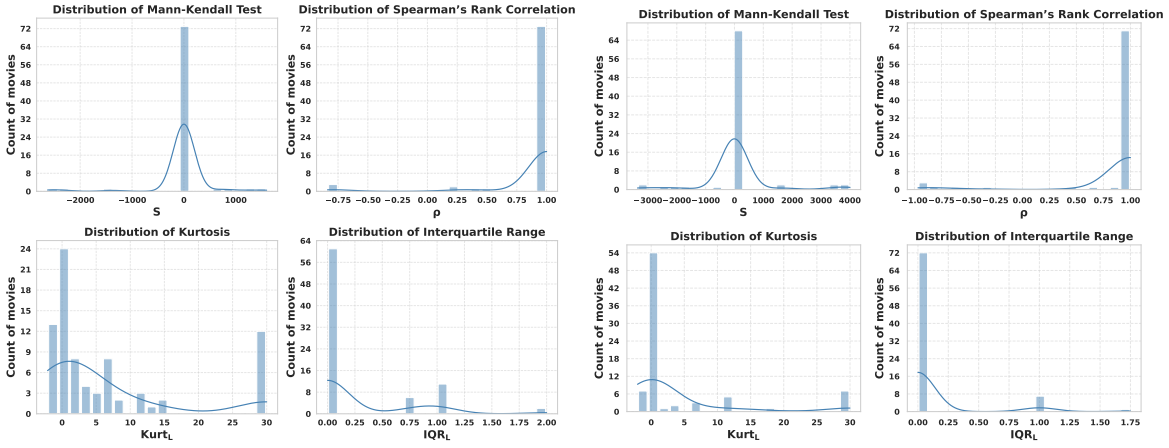



(c) "w/ Persona & w/o History"




(d) "w/ Persona & w/ History"

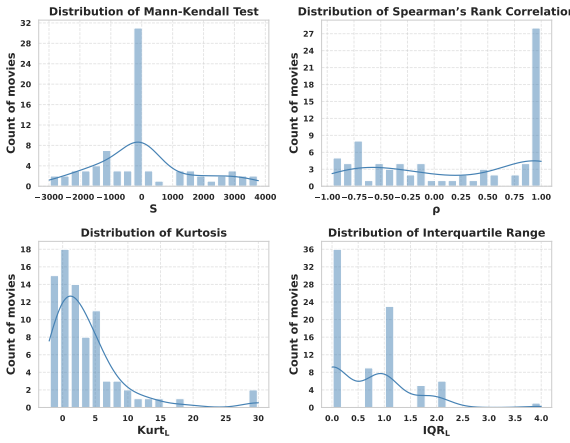
Figure 15: Distributions of Mann–Kendall Statistic, Spearman Rank Correlation, Kurtosis, Inter-quartile Range for All Movie Rating Sequences on DeepSeek-V2-Lite-Chat.



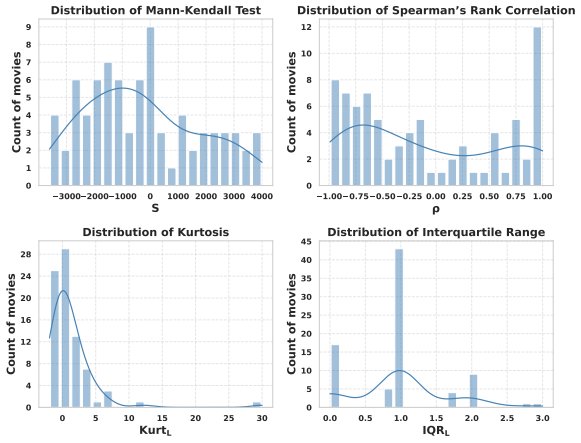

(a) "w/o Persona & w/o History"




(b) "w/o Persona & w/ History"

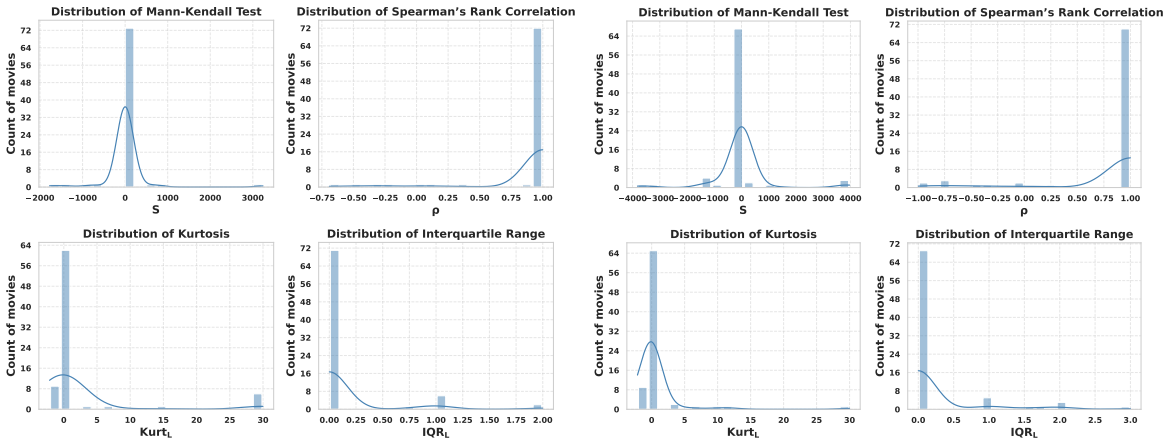



(c) "w/ Persona & w/o History"




(d) "w/ Persona & w/ History"

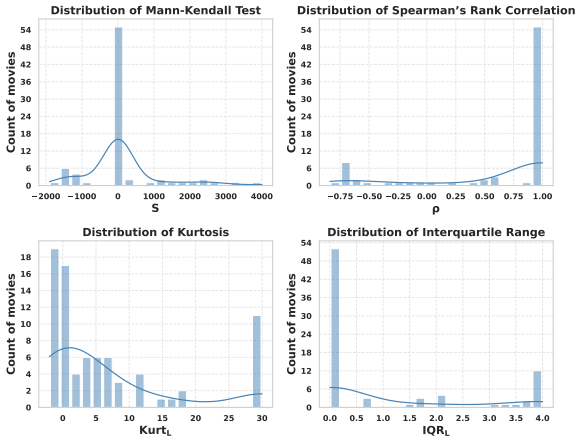
Figure 16: Distributions of Mann–Kendall Statistic, Spearman Rank Correlation, Kurtosis, Inter-quartile Range for All Movie Rating Sequences on Qwen2.5-1.5B-Instruct.



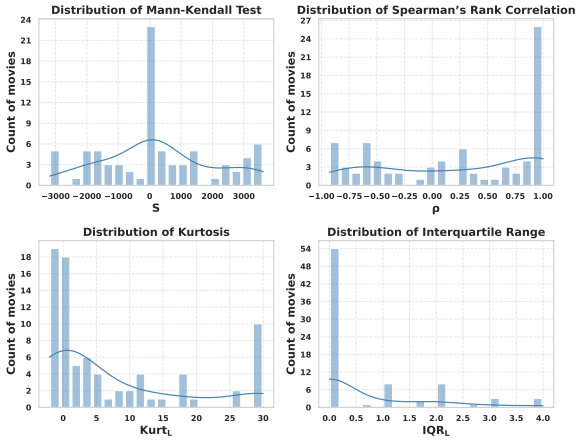

(a) "w/o Persona & w/o History"




(b) "w/o Persona & w/ History"

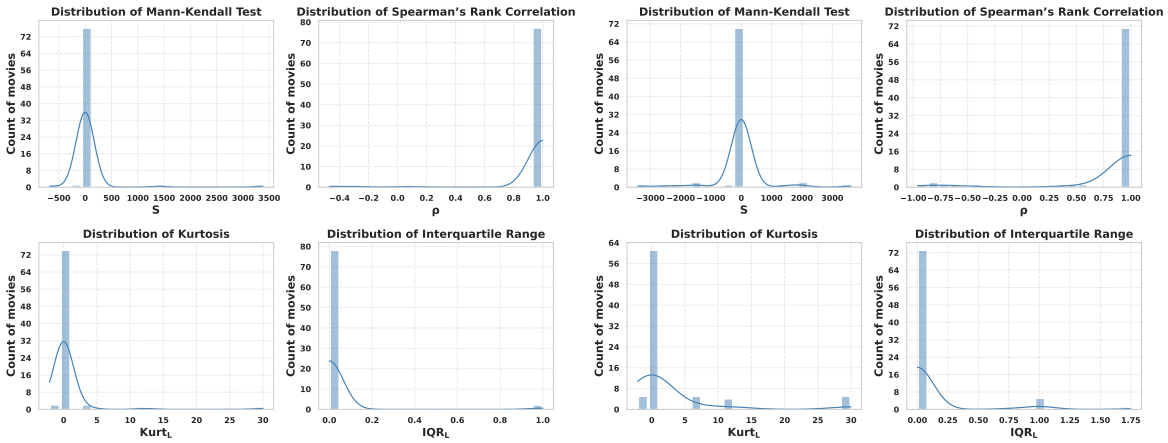



(c) "w/ Persona & w/o History"




(d) "w/ Persona & w/ History"

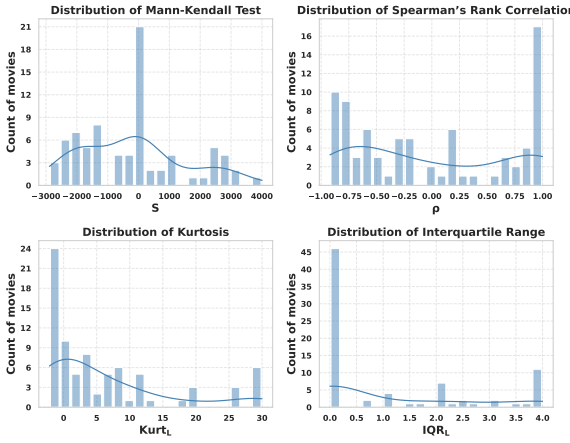
Figure 17: Distributions of Mann–Kendall Statistic, Spearman Rank Correlation, Kurtosis, Inter-quartile Range for All Movie Rating Sequences on Qwen2.5-3B-Instruct.



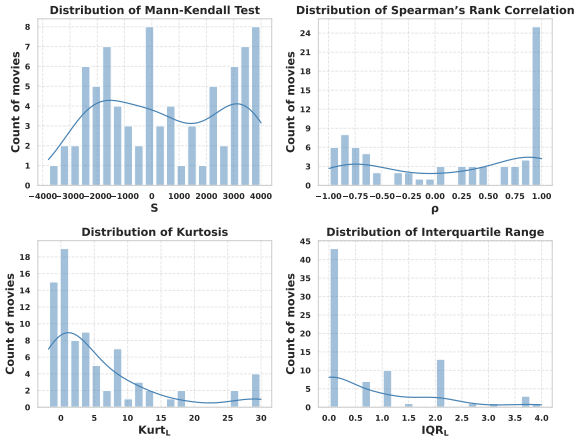

(a) "w/o Persona & w/o History"




(b) "w/o Persona & w/ History"




(c) "w/ Persona & w/o History"




(d) "w/ Persona & w/ History"

Figure 18: Distributions of Mann–Kendall Statistic, Spearman Rank Correlation, Kurtosis, Inter-quartile Range for All Movie Rating Sequences on Qwen2.5-7B-Instruct.