Spiral of Silence in Large Language Model Agents

Anonymous ACL submission

Abstract

The Spiral of Silence (SoS) theory posits that, in human societies, fear of social isolation drives individuals holding a minority opinion to quieten down, allowing the majority opinion to dominate public discourse. When agents are large language models (LLMs) rather than humans, the classic affective explanation no longer applies because language models do not have emotions or social anxiety. Therefore, a fundamental question appears: Can purely statistical language generation mechanisms give rise to SOS dynamics in collectives of LLM agents? We introduce an evaluation framework based on rating sequences and design four controlled experimental conditions by varying the presence of persona configurations and historical interaction signals. To measure opinion dynamics, we employ concentration metrics, including Interquartile Range and Kurtosis, along with trend analysis methods such as the Mann-Kendall test and Spearman rank correlation coefficient. We experiment with six widely used open-source models and a close-source model. Experimental results reveal that, in the absence of social signals, most models exhibit a strong default bias. Introducing persona configurations leads to greater opinion diversity, whereas historical collective opinion serves as an anchoring mechanism. In particular, when both persona and history are present, the dominance of the majority opinion tends to emerge more frequently, even though the agents themselves lack affective capacities. These findings challenge traditional affect-based explanations of SoS and provide empirical evidence to understand and mitigate opinion convergence in LLM-based agent systems.

1 Introduction

Social psychological dynamics often govern the formation of public opinion in human societies. One well-known phenomenon is the spiral of silence (SoS) theory, which posits that individuals are less likely to voice an opinion they perceive as unpopular due to fear of social isolation (Noelle-Neumann, 1974). As more people with minority opinions choose silence, the dominant opinion appears even more prevalent, further discouraging dissent and creating a self-reinforcing cycle in which one opinion increasingly dominates.

In contrast, agents powered by LLMs lack human emotions and social needs. LLM agents such as GPT-4 (AI, 2023) or LLaMA (Touvron et al., 2023) generate responses based on learned statistical correlations in text, not out of fear of exclusion or desire for approval. Intuitively, one might expect that a collection of LLM agents would not reproduce human-like conformity or self-reinforcing dynamics, since they have no internal concept of social isolation.

Nevertheless, some studies suggest that even without emotions, LLMs may align their outputs with trends or cues present in their input. Recent research has shown that LLM-based agent assistants sometimes tailor their responses to match the user's stated opinion (Sharma et al., 2023), and interacting LLM agents may converge toward shared conventions through repeated communication (Ashery et al., 2024). It is thus unclear whether an effect analogous to the SoS could emerge in purely LLM agent collectives, and if so, what would drive it in the absence of human-like fear of isolation.

Could LLM agents exhibit opinion convergence simply by responding to each other's outputs or to inferred majority signals, thus suppressing divergent responses despite lack of social fear? This question lies at the intersection of social psychology and LLM-based systems. Understanding whether SoS dynamics can emerge among LLM agents would reveal the extent to which complex social phenomena may arise purely from statistical language generation.

To investigate whether populations of LLM agents exhibit SoS opinion convergence in response to collective signals, we simulate a multi-agent environment. Inspired by traditional rating platforms, we design a virtual movie rating task: Many LLM agents assign ratings to movies, where ratings are considered positive and negative. We regard the historical average rating of each round as the current "collective majority opinion" signal and track how the ratio and trend of positive versus negative opinions evolve over multiple consecutive interaction rounds. In our work, LLMs respond to the prompts under different controlled conditions. We employed two key signals inspired by the SoS scenario:(1) A distinct initial persona gives each agent to describe its background and style, and (2) historical average rating summarizes the "collective majority opinion" of past in each round. By crossing these two signals, we obtain four experimental settings that allow us to isolate the effect of individual predispositions versus collective influence. Agents with personas simulate a diverse population with varying inherent initial opinions, and the historical average rating gives agents a collective majority opinion similar to a kind of social pressure.

Based on this setup, we hypothesize that if the SoS effect can emerge in LLM agents, they will increasingly conform to the perceived collective majority opinion when both historical ratings and varying initial predispositions are present. To quantify convergence, we employ two complementary classes of metrics: concentration statistics (Interquartile Range, Kurtosis) and trend diagnostics (the Mann–Kendall test, Spearman ρ). We mainstream open-source and closed-source LLMs: *DeepSeek-V2-Lite-Chat* (Liu et al., 2024), *Llama-3.1-8B-Instruct* (Grattafiori et al., 2024), *Mistral-8B-Instruct-2410*, and *Qwen-2.5-Instruct series (1.5B, 3B, 7B)* (Bai et al., 2023), covering cross-family comparisons on a similar scale and within-family scaling analyses for Qwen, and a close source model *GPT-4o-mini* (Hurst et al., 2024). The results show that in the absence of social signals, LLM agents default to positive movie ratings. Introducing a persona encourages opinion heterogeneity, while historical average ratings exert an anchoring influence. The SoS effect is significantly more likely to emerge when both signals are provided.

Our contributions of this work are as follows:

- We explored SoS theory in a measurable framework for LLM agents, introducing concentration and trend metrics to quantify the convergence of SoS.
- We conducted extensive experiments on mainstream open-source and closed-source LLMs. Our analysis covers both the performance of individual models under different settings and comparisons across model families—contrasting similarly sized models between families and differently scaled models within the same family.
- We highlighted the design and governance challenges posed by emerging opinion convergence in LLM agent systems.

2 Approach

2.1 Problem Setup

We simulate a population of LLM agents tasked with rating movies under controlled conditions to examine SoS dynamics (Noelle-Neumann, 1974). Consider an online rating system with a set of LLM agents and set of items. Agents evaluate the quality of items based on an $M \in \mathbb{N}_+$ level cardinal rating metric denoted by $\mathcal{M} \overset{\triangle}{=} \{1, \dots, M\}$, where $M \in \mathbb{N}_+$. A higher item rating implies a agent is more satisfied with an item. For example, the rating metric is $\mathcal{M} \overset{\triangle}{=} \{1 =$ "`Awful/Abysmal`", $\cdots$, 5 = "`Mediocre/Unsure`", $\cdots$, 10 = "`Perfect/Masterpiece`"}. The rating given by agent i to item j is denoted by $r_{i,j} \in \mathcal{M}$. A a gen t will give a higher rating if the item is more satisfied.

Collective opinion. Let $r_{j,k} \in \mathcal{M}$ denote the k-th observed rating of item j, where $k \in \mathbb{N}_+$. let $\mathcal{H}_{j,k}$ denote a set of all historical ratings of item j up to the k-th rating, formally:

$$\mathcal{H}_{j,k} \overset{\triangle}{=} \{r_{j,1}, \dots, r_{j,k}\}, \qquad \forall k \in \mathbb{N}_+.$$

Now, we model the formation of collective opinions, that is, the function $\mathcal{F}(\mathcal{H}_{j,k})$. agents may form a climate of opinions from the aggregation of historical ratings. Denote the collective opinion summarized from the rating history $\mathcal{H}_{j,k}$ as

$$\boldsymbol{h}_{j,k} \overset{\triangle}{=} [h_{i,k,1}, \dots, h_{j,k,M}],$$

where $h_{j,k,m} \in [0, 1]$ and $\sum_{m \in \mathcal{M}} h_{j,k,m} = 1$. The $\boldsymbol{h}_{j,k}$ is public to all agents. We consider a *class* of weighted aggregation rules to summarize historical ratings:

$$h_{j,k,m} = \frac{\sum_{l=1}^{k} \alpha_j \mathbb{I}_{\{r_{j,l}=m\}}}{\sum_{l=1}^{k} \alpha_l}, \quad \forall m \in \mathcal{M}, k \in \mathbb{N}_+ \tag{1}$$

where $\alpha_l \in \mathbb{R}_+$ denotes the weight associated with k-th rating, and $\mathbb{I}$ is an indicator function. For example, $\alpha_j = 1, \forall j$, is deployed in many rating systems, which corresponds to "*average rating rule*". Under this average rating rule, we have $h_{j,k,m} = \sum_{l=1}^{k} \mathbb{I}_{\{R_{j,l}=m\}}/k$, which is the fraction of historical ratings equal to m. Note that $\boldsymbol{h}_{j,k}$ is displayed to all agents. We capture the collective opinion formation as follows:

$$\mathcal{F}(\mathcal{H}_{j,k}) = \sum_{m \in \mathcal{M}} m \cdot h_{j,k,m} \qquad (2)$$

Persona. To emulate realistic agent diversity, each agent is assigned a different persona. Each persona is characterized by brief descriptions of occupations, backgrounds, and interests, allowing agents to reflect various predispositions and value judgments during the rating process. An example is shown in Fig. A.1.

2.2 LLM Agents Design

We design four prompts as shown in the Appendix A.2 conditions by crossing two binary signals: the presence or absence of a persona, and the presence or absence of historical average rating. All variants of prompts share a common structure: they present the information of the movie (title, genres, overview) followed by standardized instructions defining the rating scale and requesting an output. In all cases, agents are instructed to output only a single rating from 1–10 based on a brief scale. The four conditions are as follows:

- **w/o Persona & w/o History:** The baseline is designed by only the movie information without the current historical average rating and the rating instructions. The agent has no persona context. This setting represents an independent agent with no social influence.
- **w/ Persona & w/o History.** With a persona description for the agent. This gives the agent a fixed identity or background knowledge, but no social influence from others' opinions. Simulate an agent with persona acting independently.
- **w/o Persona & w/ History:** The agent without persona, but the movie information includes the current historical average rating of the movie (on the scale of 1-10) based on the ratings of previous agents.
- **w/ Persona & w/ History:** Combine both signals, the agent is assigned a persona and shown the current historical average rating from previous agents while meanwhile. This condition models a persona-driven agent under social influence, where the agent's intrinsic preferences (via persona) may interact with pressure to conform to the collective opinion displayed.

Under "w/ history" conditions, the "historical average rating" is updated in real time as the rating progresses. By comparing these four setups, we can disentangle the effect of an agent's persona from the effect of seeing others' opinions on the emergence of consensus or the silencing of minority opinion.

2.3 Rating Procedure

Each movie rating task is designed as a sequential rating process by multiple independent agents for the same movie, to observe how dominant opinions evolve over multiple rounds. For a given movie, agents take turns providing a rating one after another in a random order. If persona are in use, we randomly sample 100 unique persona from the PersonaHub subset for that movie, assigning a different persona to each agent to ensure diversity of backgrounds. If persona is not used, each agent is effectively identical, but we still treat each rating act as a separate agent instance.

Sequential rating update: In the sitting "w/ History", within a movie's 100 agent sequence, agents take action one by one in all rounds, the prompt for agent n includes the average ratings of all agents from 1 to $n-1$ (For the first agent, we initialize the "current historical average rating" using the IMDb average rating as an initial public collective opinion). After agent n produces a rating, we update the historical average rating to include that new rating, and then agent $n+1$ could see this updated collective opinion.

Multi-sample stable evaluation: Each agent's rating is obtained by averaging three independent model outputs from identical prompts, then rounding to the nearest integer. This reduces stochastic variance and yields a more stable estimate of the agent's opinion. The process is repeated for all 100 agents per movie. Each model follows this procedure across 80 movies under four experimental settings.

2.4 How to quantify?

Our goal is to detect whether repeated sequential ratings generated by LLM agents display SoS effects: minority opinion fades, while the majority opinion becomes increasingly dominant. For movie j we obtain a sequence of $T = 100$ integer ratings $\mathcal{H}_{j,T} = \{r_{i,j,k}\}_{k=1}^{T}$, where $i = i(j,k)$ indexes the agent "w/ Persona" that produced the k-th rating.[1] Let $\mathcal{H}_{j,k} = \{r_{i,j,1}, \ldots, r_{i,j,k}\}$ denote the first k ratings for movie j and let $\mathcal{F}(\mathcal{H}_{j,k}) = \frac{1}{k}\sum_{t=1}^{k} r_{i,j,t}$ be the historical average rating.

Rating distance. For every individual rating we measure its deviation from the current collective opinion (historical average rating):

$$\mathrm{Dist}(r_{i,j,k}) = \big| r_{i,j,k} - \mathcal{F}(\mathcal{H}_{j,k}) \big|. \qquad (3)$$

Majority-opinion trend. Let $\mathrm{pos}_{j,k}$ and $\mathrm{neg}_{j,k}$ denote the cumulative proportions of positive and negative ratings for movie j up to step k, respectively. They are defined as:

$$\mathrm{pos}_{j,k} = \frac{1}{k}\sum_{t \le k} \mathbf{1}\big[r_{i,j,t} \ge 6\big],$$
$$\mathrm{neg}_{j,k} = \frac{1}{k}\sum_{t \le k} \mathbf{1}\big[r_{i,j,t} \le 5\big].$$

where $\mathbf{1}[\cdot]$ is the indicator. We define the textitdynamic majority-conforming opinion (MCO) sequence as:

$$\mathrm{MCO}_{j,k} = \max\big\{\mathrm{pos}_{j,k},\, \mathrm{neg}_{j,k}\big\}, \quad k = 1, \ldots, T.$$

To reduce the impact of early-stage fluctuations, all the following trend statistics are computed starting from round $m = 11$:

- **Mann–Kendall Statistic** (S) is used to detect monotonic trends in the majority choice series (Mann, 1945). $S_j = \sum_{k=m}^{T-1}\sum_{t=k+1}^{T} \mathrm{sgn}\big(\mathrm{MCO}_{j,t}\mathrm{MCO}_{j,k}\big)$, where $S_j{>}0$ indicates a monotone increase in majority support.
- **Spearman Rank Correlation** (ρ) is used to quantify rank-based reinforcement of the majority choice (Spearman, 1904). The Spearman ρ_j between the index $k = m{:}T$ and $\mathrm{MCO}_{j,k}$ captures the strength of the upward trend; $\rho_j \to 1$ implies nearly perfect reinforcement.

Rating concentration. If SoS develops, the late portion of the rating sequence should be tight around a majority value. We measure dispersion on the final L rounds ($L = 30$ by default) with two complementary statistics:

- **Inter-quartile Range (IQR)** quantifies the central spread of recent ratings, providing a robust dispersion measure (Clark-Carter, 2005). $\mathrm{IQR}_L(j) = Q_3 - Q_1$, where Q_1 and Q_3 are the 25th and 75th percentiles of the last L ratings $\mathcal{H}_{j,T-L:T}$ for movie j.
- **Kurtosis** describes the tailedness of the rating distribution, highlighting the presence of outliers (Balanda and MacGillivray, 1988). $\mathrm{kurt}_L(j) = \frac{1}{L}\sum_{k=T-L+1}^{T}\left(\frac{r_{i,j,k}-\mu_j}{\sigma_j}\right)^4 - 3$, where μ_j and σ_j are the mean and standard deviation of the same L ratings.

Smaller IQR_L together with positive kurt_L signals a sharp, concentrated majority.

Decision thresholds. A movie's rating sequence is marked **True** (i.e., "Spiral") if all of the following hold:[2]

- $\mathrm{MCO}_{j,T} \ge 0.65$, the terminal majority opinion comprises at least 65% of all ratings;
- $S_j \ge 50$ or $\rho_j \ge 0.60$, the series exhibits a statistically salient upward trend, reflected either by a slope of at least 50 or a Spearman correlation no less than 0.60;
- $\mathrm{kurt}_L(j) > 0$ or $\mathrm{IQR}_L(j) < 2$, late-stage ratings are tightly clustered, indicated by positive kurtosis or an inter-quartile range below 2.

Semantic match between persona and movie overview. When personas are available, we examine how the semantic match between a user's persona and a movie influences deviations from the collective rating. Specifically, for each agent i and movie j, we compute a semantic match score, defined as the cosine similarity $\mathrm{sim}(i,j)$ between the TF–IDF embedding of the agent's persona and the movie overview. Each rating yields a data point $\big(\mathrm{sim}(i,j),\, \mathrm{Dist}(r_{i,j,k})\big)$, where Dist denotes the absolute deviation from the historical average rating for that movie. We analyze the correlation between semantic match and rating deviation across all agent–movie instances.

3 Experiments

We investigate whether a SoS effect emerges in LLM agents by rating movies under four configurations: (1) w/o Persona & w/o History, (2) w/ Persona & w/o History, (3) w/o Persona & w/ History, and (4) w/ Persona & w/ History. In each setting,

[1] Since each agent rates a movie at most once, i is uniquely determined by (j,k).

[2] Thresholds were fixed *a priori*, then applied to every model/condition.

we analyze the rating sequence for 80 movies using both qualitative visualizations and statistical measures. We focus primarily on `GPT-4o-mini` as a case study model. To quantify the "spiral" formation, we track the majority opinion fraction MCOj, k for movie j after k rating rounds. We also calculate the interquartile range (IQR) and kurtosis (to measure the dispersion and peakedness of the rating distribution) of the final rating sequence, as well as the Mann–Kendall (MK) statistic and Spearman's ρ to detect monotonic rating trends over rounds.

3.1 Setup

We evaluate a diverse set of LLM agents, including `DeepSeek-V2-Lite-Chat`, `Llama-3.1-8B-Instruct`, `Mistral-8B-Instruct-2410`, `Qwen-2.5 series (1.5B, 3B, and 7B)`, and `GPT-4o-mini`. For each combination of model and condition, we randomly sampled 80 movies from our dataset to serve as rating tasks. This yields a broad evaluation across different content. All generations were performed with a fixed temperature of 0.1 to reduce randomness and improve reproducibility. Experiments are conducted on 4×NVIDIA A100 GPUs.

Each agent provides a movie rating on a 10-point integer scale (1 = `Awful/Abysmal`, 5 = `Mediocre/Unsure`, 10 = `Perfect/Masterpiece`). For conceptual clarity, we treat the ratings ≥ 6 as a positive (favorable) opinion and ≤ 5 as negative (unfavorable). This binary split allows us to later analyze majority vs. minority opinion formation. We drew on two key datasets: personas and movies. Agent personas are sampled from the top 10,000 entries of the *elite_persona*[3] subset from the PersonaHub dataset (Ge et al., 2024), which provides a textual profile and domain for each persona (for example, a brief self-description or background). These personas served as static agent attributes that can influence an agent's initial preferences.

For movie data, we compiled a dataset of films released after January 12, 2025 by scraping IMDb[4]. This cutoff date ensured that the movies are unlikely to appear in the training data of our models (preventing any memorization or prior knowledge). For each movie, we collected its title, genres, a brief overview (synopsis), the IMDb's average rating (as of scraping time), and the number of IMDb user ratings. We filtered out movies with fewer than 30 IMDb ratings to focus on films with an established baseline of public opinion.

3.2 Scenario I: w/o Persona & w/o History - Prior Positivity Bias

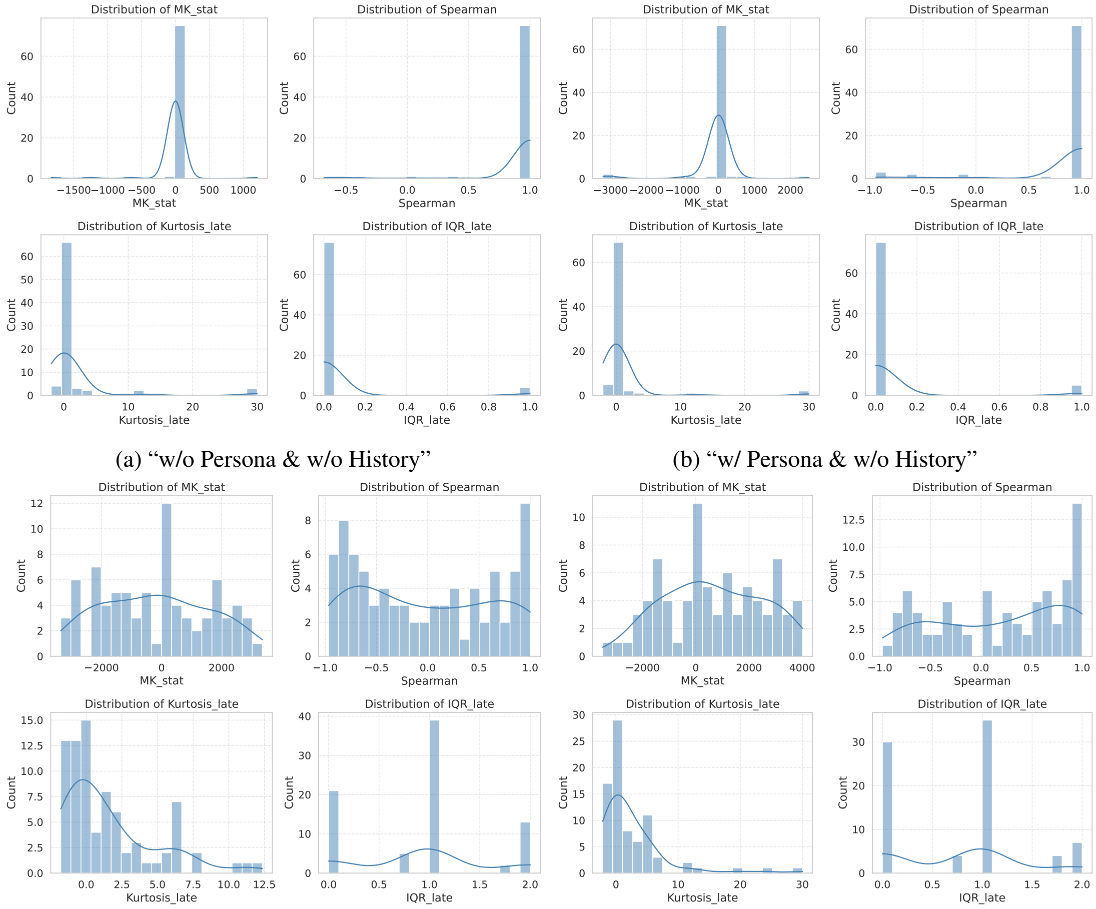
In this condition, see Fig.10 in Appendix A.4, the trend plots of the opinion proportion for each movie (where the uppermost line is equivalent to the MCO curve) reveal a highly static and extreme state: for almost all samples, the positive opinion proportion (orange line) keeps to 1.0 from the first round onward, while the negative opinion proportion (blue line) hovers near 0. In addition, as shown in Fig.1 (a), Mann-Kendall statistic and Spearman concentrate at the extreme positive end, indicating a strongly monotonic sequence with no reversals. Kurtosis is mostly positive and IQR is extremely low, suggesting that the ratings are tightly clustered with minimal divergence at the final round rating. Taken together, these trends and statistics suggest that the "majority opinion" in this setting is not the result of genuine group evolution or social influence, but rather reflects a "positivity prior" embedded in the model training data. In the absence of agent diversity or historical anchors, the model output defaults to a static and monolithic majority.

3.3 Scenario II: w/ Persona & w/o History - Anchoring and Adjustment

As shown in Fig.11 in Appendix A.4, the opinion proportion trend plots (MCO curves) show a different but equally extreme phenomenon, the positive or negative (orange or blue line) randomly dominates at the start of the sequence, then remains highly stable throughout all rounds, with almost no cross-over or reversal. That is, the majority opinion is anchored by the historical average rating, without further adjustment. We can see from Fig.1 (b), the corresponding trend statistics metrics Mann_Kendall and Spearman are similarly skewed toward the positive extreme, indicating that once a "dominance" side is established, there is no further trend shift. IQR and Kurtosis concentration metrics show that late-stage ratings are highly concentrated. Notably, a small minority of movies show rare jumps (occasional crossovers between positive and negative lines), but such cases are exceedingly rare. This behavior exemplifies the classic "Anchoring and Adjustment" cognitive bias: The historical average serves as an anchor for the initial rating, and all subsequent ratings are minor adjustments around this anchor. In the absence

[3] https://huggingface.co/datasets/proj-persona/PersonaHub

[4] https://www.imdb.com/search/title/?title_type=feature&release_date=2025-01-12,2025-12-31&sort=release_date,asc



(a) “w/o Persona & w/o History”

(b) “w/ Persona & w/o History”

(c) “w/o Persona & w/ History”

(d) “w/ Persona & w/ History”

Figure 1: Distributions of Mann–Kendall Statistic, Spearman Rank Correlation, Kurtosis, Inter-quartile Range for All Movie Rating Sequences on GPT-4o-mini.

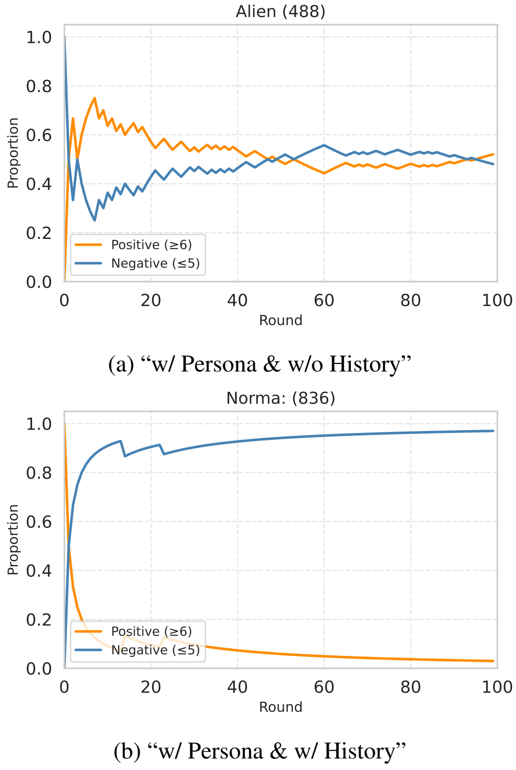
of individual differences or external shocks, the model automatically locks onto the anchor point, severely dampening the conditions necessary for a or dynamic opinion change.

3.4 Scenario III: w/o Persona & w/ History - Competing Perspectives

In Fig.2 (a), a case shows that the majority-conforming opinion (MCO) curves for the movies exhibit marked fluctuations: both the positive (orange line) and negative (blue line) opinion proportion curves oscillate persistently throughout the rating rounds, with neither side achieving sustained dominance. As shown in Fig.1 (c), the distribution of trend and concentration metrics further supports this observation. The distributions of Mann-Kendall statistic and Spearman are centered near zero, indicating little to no overall monotonic trend. Likewise, the Kurtosis tends to be close to zero or negative, and the interquartile range (IQR) remains moderate to high, reflecting a relatively flat or dispersed final rating distribution. This result suggests that the introduction of diverse agent personas, in the absence of historical average rating as anchor, injects substantial heterogeneity into the system. Each agent rates the same movie from distinct internal and initial perspectives, leading to persistent competition between the majority opinions. Under these circumstances, the formation of the SoS is rare; the emergence of majority dominance might require additional socially influential mechanisms.

3.5 Scenario IV: w/ Persona & w/ History - Emergence of Spiral of Silence

As shown in Fig.2 (a), in this setting, the trend of the positive / negative opinion proportions for typical movies displays a clear pattern of convergence of the majority opinion. Regardless of the initial disagreement, once one side gains an advantage in the early rounds, its dominance continues to strengthen and quickly becomes stable, suppressing the minority. This echoes the : the majority collective opinion becomes increasingly



(a) “w/ Persona & w/o History”

(b) “w/ Persona & w/ History”

Figure 2: Cases of GPT-4o-mini.

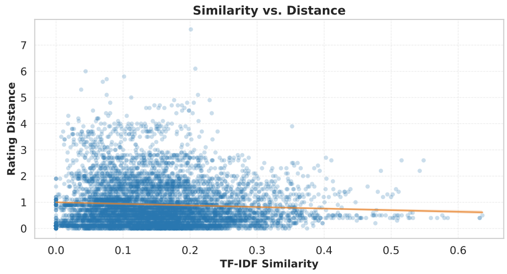


Figure 3: The relation between **Semantic Similarity vs. Rating Distance**.

entrenched, while dissenting voices silence. From the distribution of four metrics shown in Fig.1 (d), the Mann-Kendall Statistic and Spearman are predominantly positive, Kurtosis is mostly above zero, and IQR is sharply reduced. This indicates that for most movies, final ratings are highly concentrated around the majority opinion, with the convergence trend being both monotonic and statistically significant. Mechanically, two signals are at play: Persona diversity injects initial disagreement, ensuring opinion plurality among agents; the history average rating as anchor amplifies social pressure, so once a majority emerges, subsequent agents conform more easily, and dissent quickly fell silent. The result: monotonic changes in the positive/negative opinion proportion, stable majority dominance, and spontaneous emergence of the spiral.

Semantic Similarity vs. Rating Distance To further investigate the micro-mechanisms underlying the , we examine the relationship between the semantic similarity of agent personas and movie overviews (measured by TF-IDF (Salton and Buckley, 1988; Pedregosa et al., 2011) cosine similarity) and the rating distance was defined in Eq.**??** under the “w/ Persona & w/ History” condition. As shown in Fig.3, when persona–history similarity is low, large rating distances frequently occur; when similarity is high, substantial deviations are rare. Here, similarity refers to the semantic match between the agent’s persona and the movie overview. This suggests that when an agent’s persona is highly aligned with the overview of a movie, the model was much more likely to give consensus-conforming ratings, rarely deviating from the majority opinion; while low similarity increases the agent’s propensity to dissent, leading to a large rating distance even in the presence of strong social influence. This observation provides additional support for the notion that semantic consistency between persona and movie overview is a critical prerequisite for the emergence of the SoS in persona-conditioned settings.

Model	**sos (%)**
`GPT-4o-mini`	27.5
`DeepSeek-V2-Lite`	15.0
`Mistral-8B-Instruct-2410`	27.5
`Qwen-2.5-Instruct-1.5B`	11.3
`Qwen-2.5-Instruct-3B`	16.3
`Qwen-2.5-Instruct-7B`	28.8

Table 1: Proportion (%) of SoS under **w/ Persona & w/ History** setting for different model.

3.6 Cross-Model Comparing

The above findings for GPT-4o-mini raise the following question: Do other LLMs exhibit similar SoS tendencies under the same conditions? We examined under the **w/ Persona & w/ History** setting in several models and found a noticeable variation in the prevalence of SoS. In particular, we compared `GPT-4o-mini` with `DeepSeek-V2-Lite-Chat`, `Mistral-8B-Instruct-2410`, `Qwen-2.5-1.5B-Instruct`, `Qwen-2.5-3B-Instruct`, `Qwen-2.5-7B-Instruct`. All models were given identical persona prompts and rating tasks. We reported the incidence rates of the SoS under the w/ Persona & w/ History setting for different models, as shown in Table 1. **Cross-family** Compared to `DeepSeek-V2-Lite`, models such as `GPT-4o-mini`, `Mistral-8B-Instruct`, and Qwen-2.5-7B-Instruct were more likely to converge to a unified opinion, indicating that DeepSeek-V2-Lite is less prone to fully suppressing dissenting views. This observation may be attributed to differences in pre-training data, training paradigm, or the degree of instruction tuning across models. Detailed results for these open source models are provided in Appendix A. **Within-family** Across the Qwen-2.5 series (1.5B to 3B to 7B), as the model size increases, both monotonicity and the strength of majority convergence are enhanced: larger models consistently exhibit higher values of the Mann-Kendall statistic and Spearman, reflecting stronger and more persistent reinforcement of majority opinions. In parallel, the concentration of the final scores increases, as indicated by lower IQR values, suggesting a stronger consensus between the agents. These trends implied that scaling up the model size amplifies both the tendency toward monotonic majority dominance and the eventual unanimity of group ratings. Detailed results for each Qwen model are provided in the Appendix. Surprisingly, the model Llama-3.1-8B-Instruct behaves quite unlike the others: it almost never exhibits SoS convergence and, even without any persona cues, frequently oscillates between opposing opinions. A detailed discussion of this unexpected pattern is provided in the Appendix A.6.

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 social media users are far more likely to voice opinions when they believe their network agrees with them, and Porten-Cheé and Eilders (2015) demonstrates that anonymity and low-effort feedback significantly increase willingness to express unpopular views in 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) studies a community of LLMs playing the social-deduction game Avalon. Akata et al. (2025) use behavioral game theory to let LLMs 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 via a probabilistic SUVA framework and reports that most models' decisions reflect social welfare and reciprocity considerations rather than pure self-interest. Other recent work focuses on how LLMs represent majority opinions (Weng et al., 2025). For example, Ye et al. (2024) systematically quantifies biases in "LLM-as-judge" scenarios and identifies a strong bandwagon effect. To explore opinion dynamics, Nasim et al. (2025) proposes a simulator that embeds LLM-based agents into networked opinion-spread models. By integrating classic social-influence theories (Kelman, 1958; Munroe, 2013) with LLM communication, their framework lets researchers 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 yield 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 show that LLM-based movie rating agents exhibit a clear positivity bias by default, yet develop a richer spectrum of 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 an 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 stand-in 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 spiral of silence 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 "Spiral of Silence". Although this advances scientific understanding, it also entails several risks: Malicious actors could adapt our protocol to build large-scale "astroturf" 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 maintainers and contributors of the open-source models DeepSeek-V2-Lite-Chat, Llama-3.1-8B-Instruct, Mistral-8B-Instruct-2410, and the Qwen-2.5 series, as well as the authors of the PersonaHub dataset for making their resources publicly available. Large-language-model tools (GPT-4 family) were used for wording suggestions and grammar checks; all content and analyses were verified by the authors, who bear sole responsibility for any remaining errors.

References

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

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*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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.

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.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

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.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

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*.

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 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 4: An Persona Example

A.2 Rating Prompts

We present the four distinct prompts employed in our experiments. Each prompt corresponds to one of the four experimental settings explored in the study, differing by the presence or absence of historical average rating information (**history**) and character profile information (**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: w/o Persona & w/o History

```
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.
}
```

Prompt: w/o Persona & w/ History

```
Please provide your rating for the movie.

# Movie Information
Title: [Movie Title]
Genres: [Genres]
Overview: [Movie Overview]
Movie average rating: [Historical Avg] (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: w/ Persona & w/o History

```
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: w/ Persona & w/ History

```
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 Avg] (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.
```

A.3 Metrics Distributions for Other Models

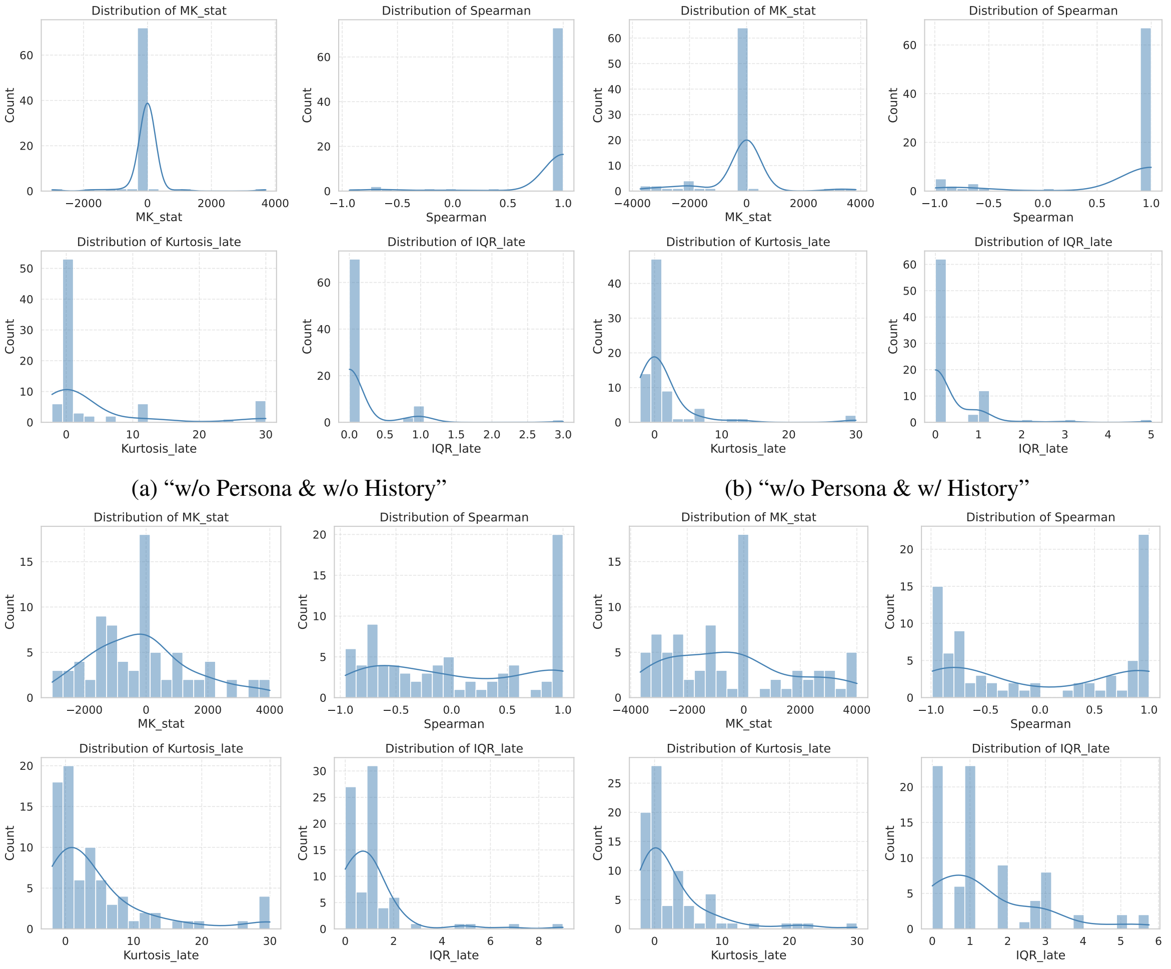
To provide a comprehensive comparison beyond the main case study (GPT-4o-mini), we report the distribution of statistical metrics for all models under each four settings. For each model, we visualize and summarize four key statistics across all movies: Mann–Kendall statistic (S), Spearman's ρ, late-stage interquartile range (IQR), and late-stage kurtosis (Kurtosis) . These results complement the qualitative and quantitative analyses in the main text and highlight model-specific tendencies regarding majority convergence, opinion monotonicity, and rating concentration.

A.3.1 Experimental Settings

For each of the following models, we conducted the movie rating under each four settings. For each model and condition, we explore the distribution of the four metrics.

A.3.2 Results for Open-Source Models

For each setting, the metrics reflect the model's tendency (or lack thereof) to form persistent majority opinions and converge to collective opinion.



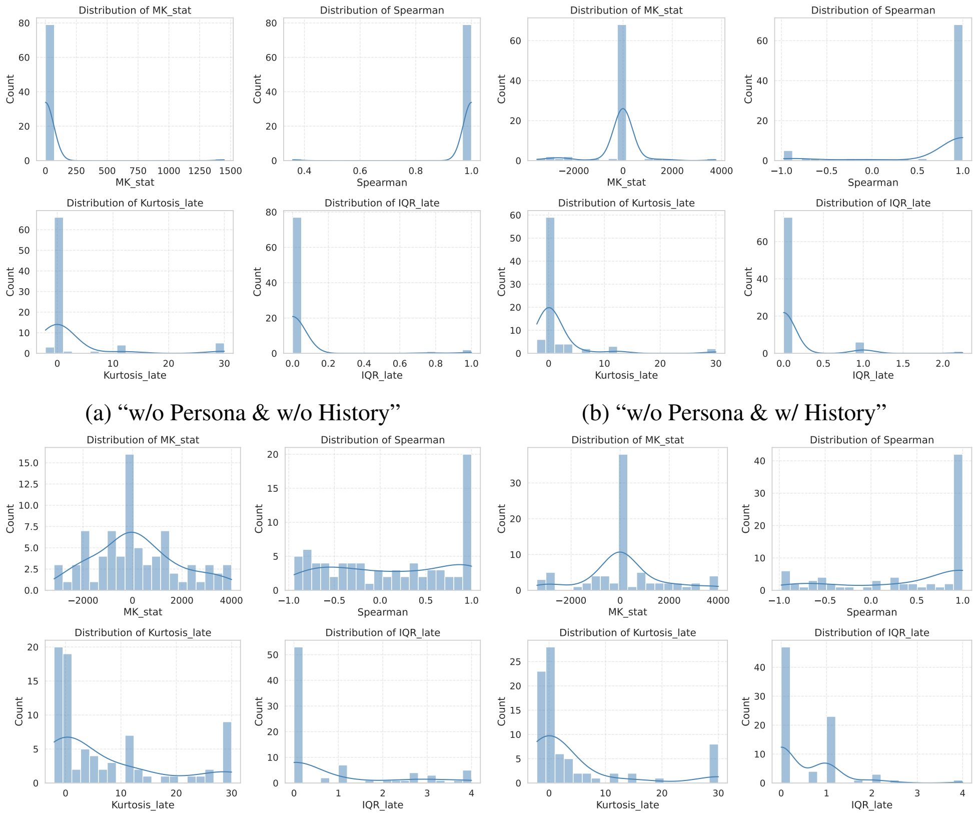
(a) “w/o Persona & w/o History”

(b) “w/o Persona & w/ History”

(c) “w/ Persona & w/o History”

(d) “w/ Persona & w/ History”

Figure 5: Distributions of Mann–Kendall Statistic, Spearman Rank Correlation, Kurtosis, Inter-quartile Range for All Movie Rating Sequences on DeepSeek-V2-Lite-Chat.



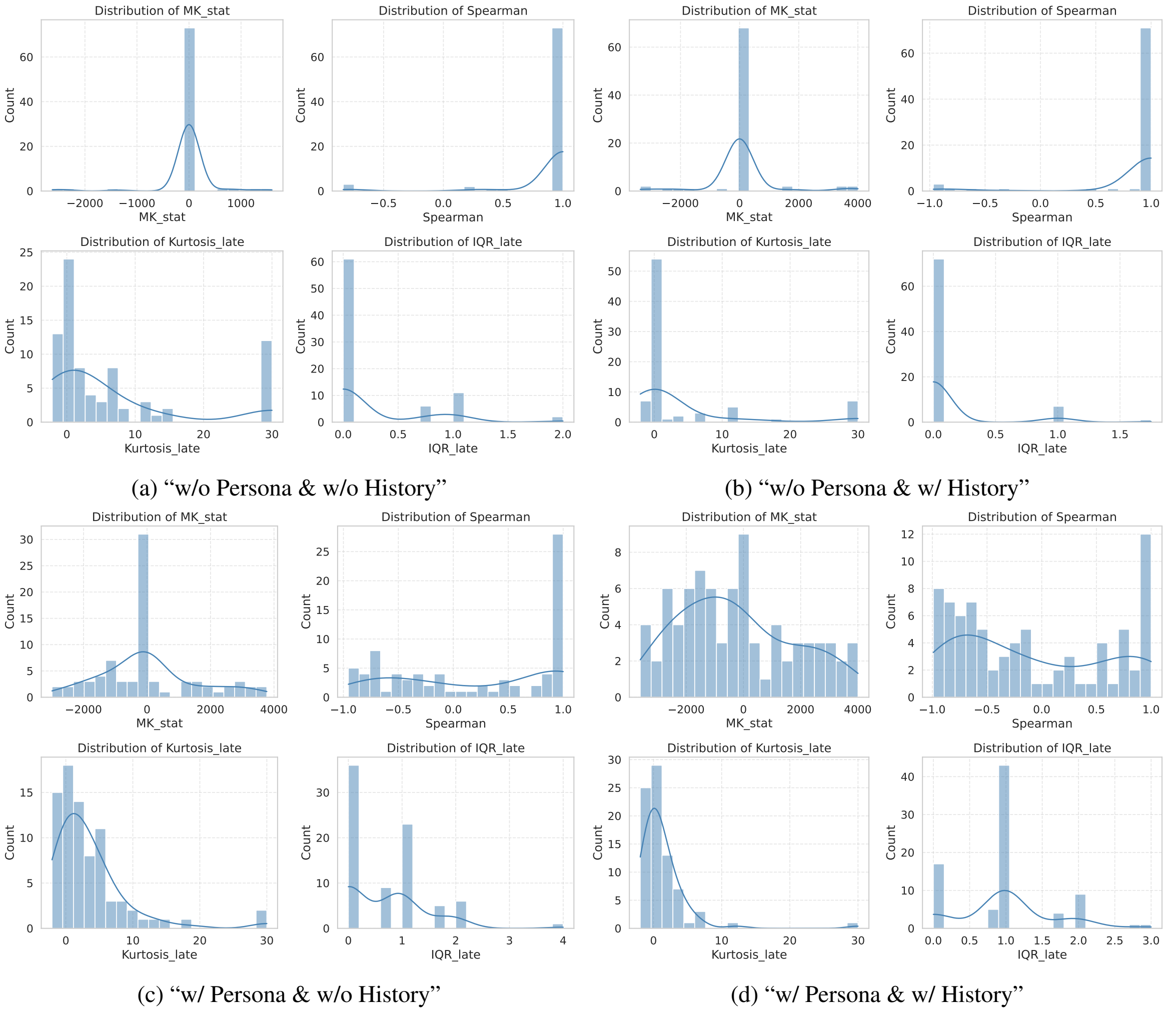
(a) “w/o Persona & w/o History”

(b) “w/o Persona & w/ History”

(c) “w/ Persona & w/o History”

(d) “w/ Persona & w/ History”

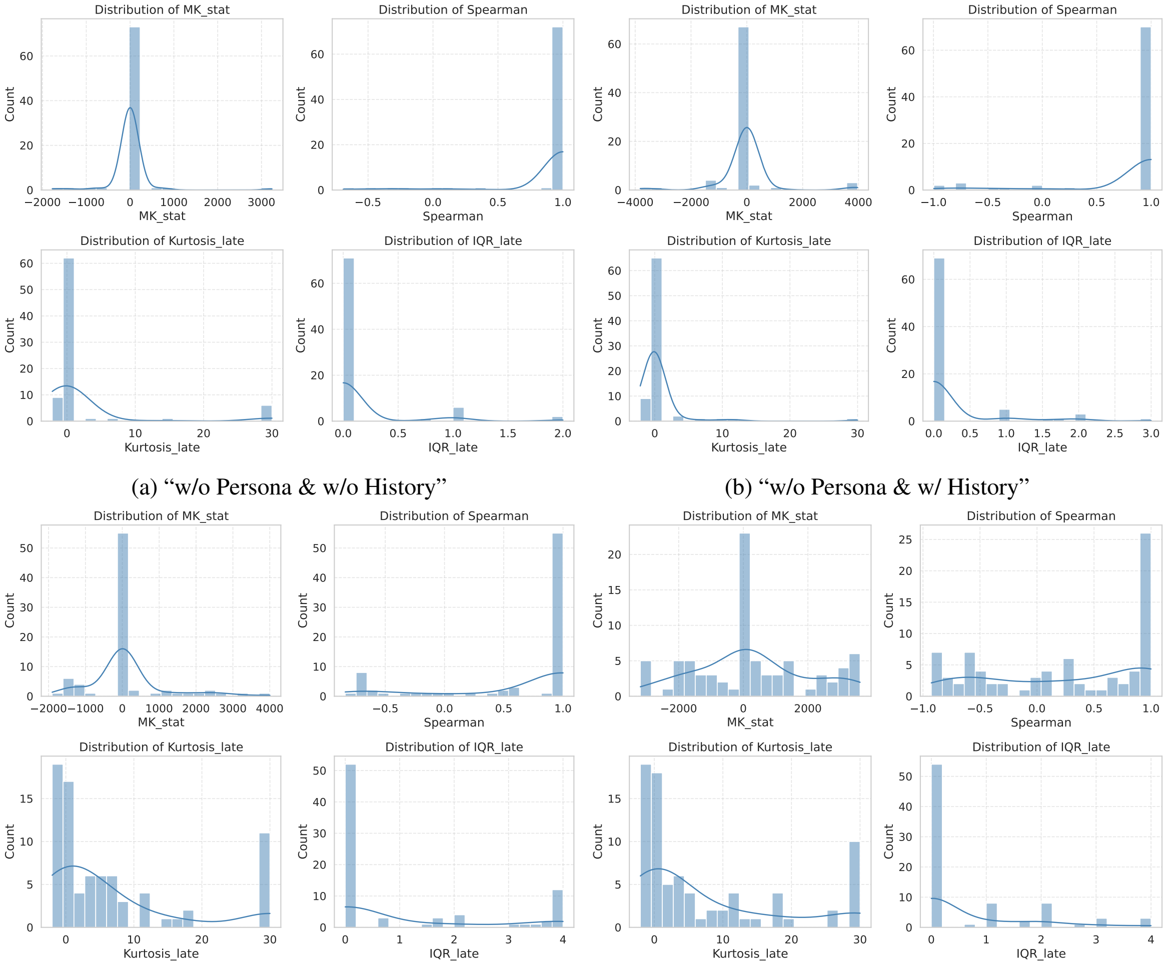
Figure 6: Distributions of Mann–Kendall Statistic, Spearman Rank Correlation, Kurtosis, Inter-quartile Range for All Movie Rating Sequences on Ministral-8B-Instruct-2410.



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

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

Figure 7: Distributions of Mann–Kendall Statistic, Spearman Rank Correlation, Kurtosis, Inter-quartile Range for All Movie Rating Sequences on Qwen2.5-1.5B-Instruct.



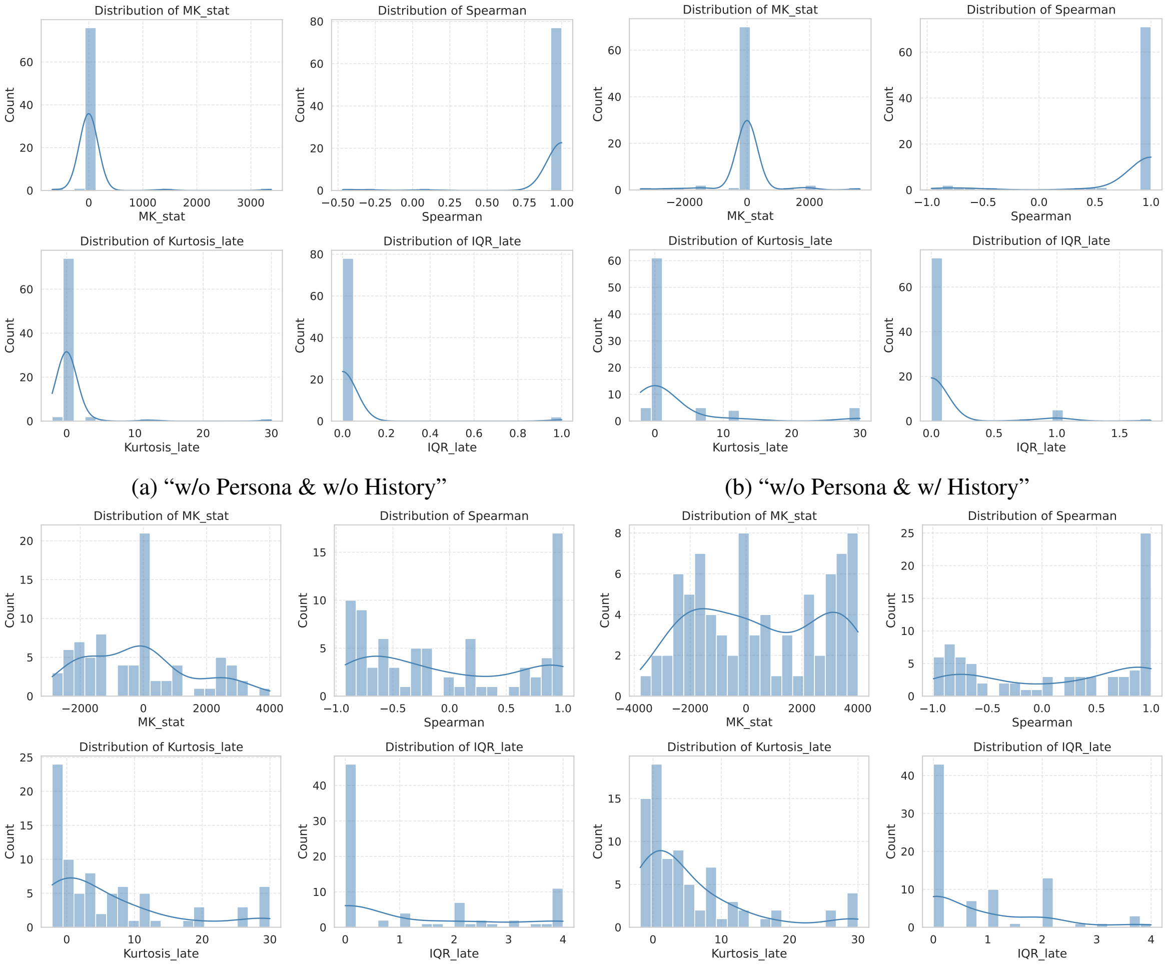
(a) “w/o Persona & w/o History”

(b) “w/o Persona & w/ History”

(c) “w/ Persona & w/o History”

(d) “w/ Persona & w/ History”

Figure 8: 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 9: Distributions of Mann–Kendall Statistic, Spearman Rank Correlation, Kurtosis, Inter-quartile Range for All Movie Rating Sequences on Qwen2.5-7B-Instruct.

A.4 Positive and Negative Opinion Proportion Trends (MCO Trend)

(**Note:** It should be emphasized that, in all visualizations, the trajectory of the opinion proportion that occupies the upper position (be it positive or negative) is, by definition, equivalent to the MCO (majority-conforming opinion) curve for that sequence.)

In this section, we present the results of positive and negative opinion proportion trends (i.e. MCO curves) for each language model under four experimental settings.

For each combination of model and setting, we include 80 trend plots (one per movie) showing the proportion of positive vs. negative opinions across 100 rating rounds. In these plots, a “positive” opinion is defined as a movie rating ≥ 6, while a “negative” opinion corresponds to a rating ≤ 5. Each subplot thus traces the fraction of positive opinions (and implicitly, negative opinions) over time for a single movie. The 80 subplots are arranged in a compact grid of 16 rows × 5 columns per page.

These visualization figures provide raw visual evidence of possible SoS emerging over time. In particular, one can observe whether initially minority opinions tend to diminish as rounds progress (which would be indicative of a SoS effect) or whether they persist.

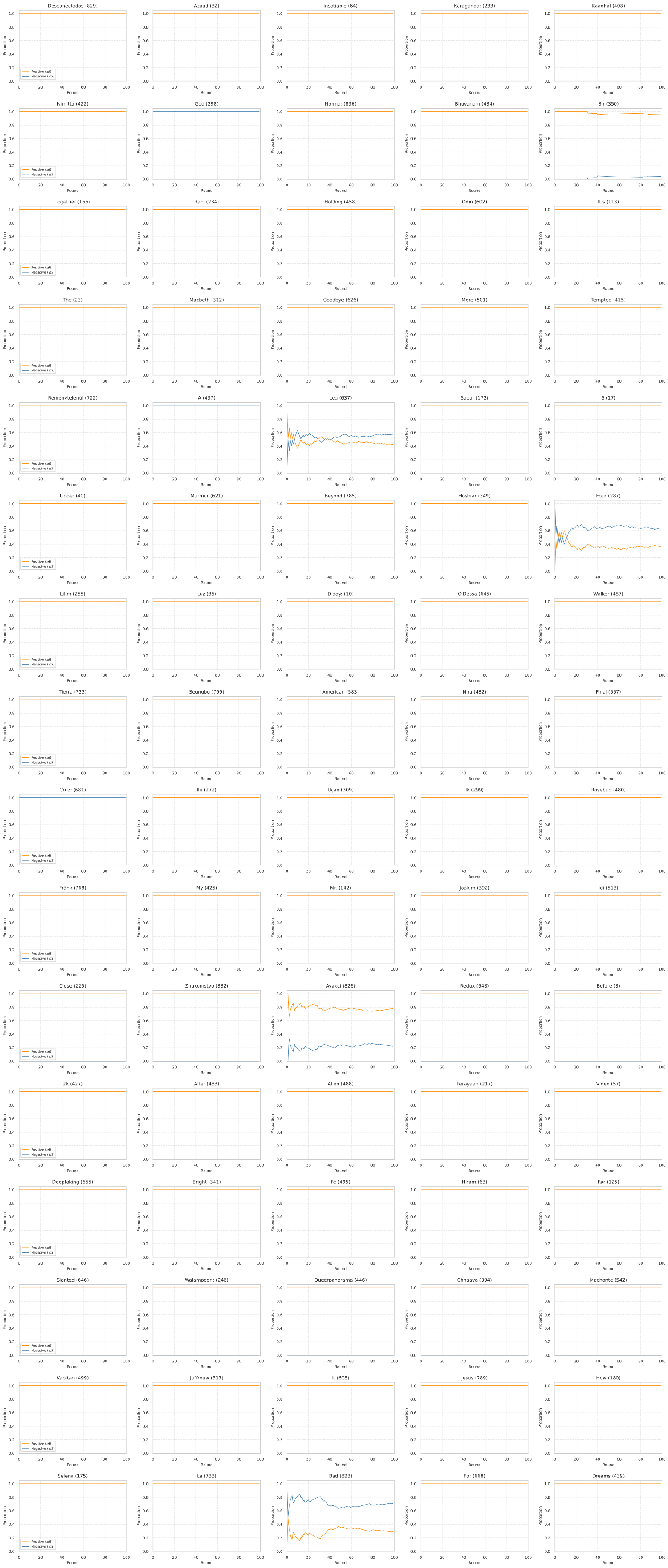


Figure 10: Trend of Positive and Negative Opinion Proportions on GPT-4o-mini under the “w/o Persona & w/o History” Setting.

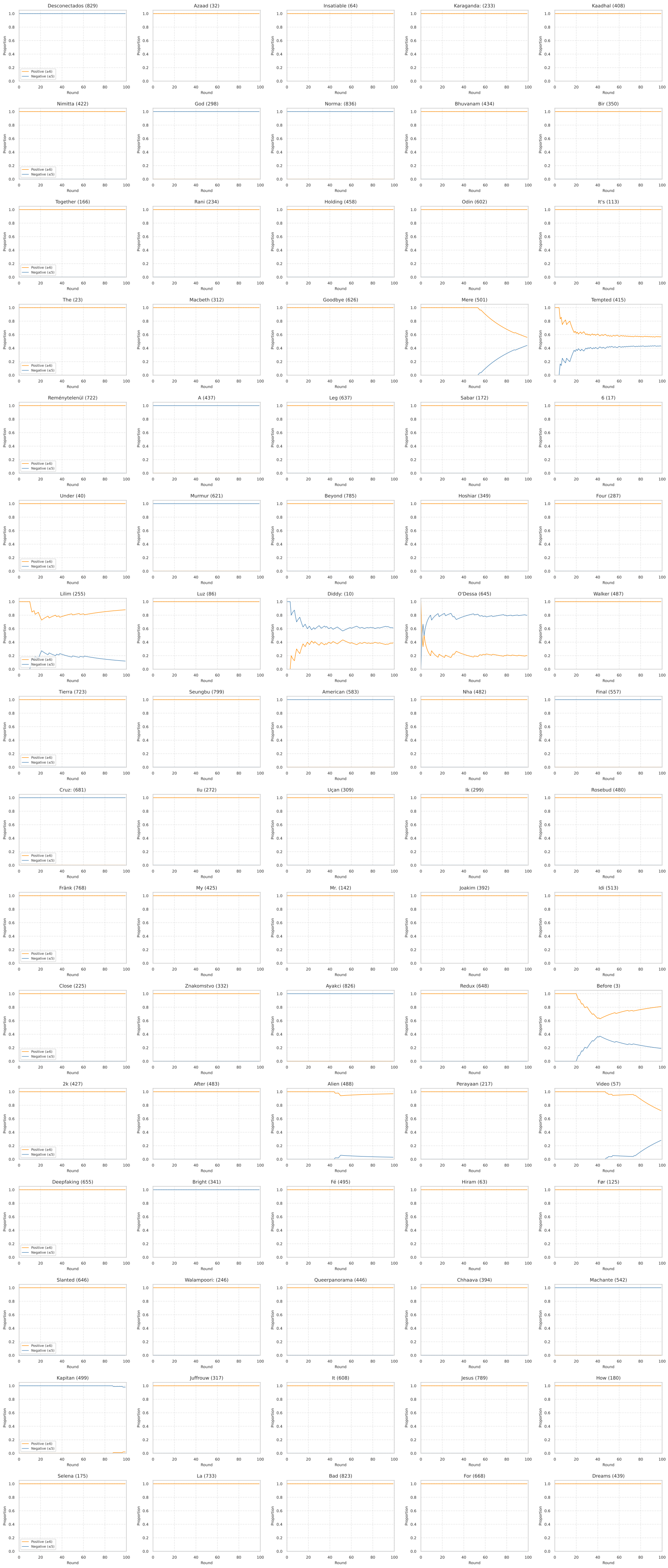


Figure 11: Trend of Positive and Negative Opinion Proportions on GPT-4o-mini under the “w/o Persona & w/ History” Setting.

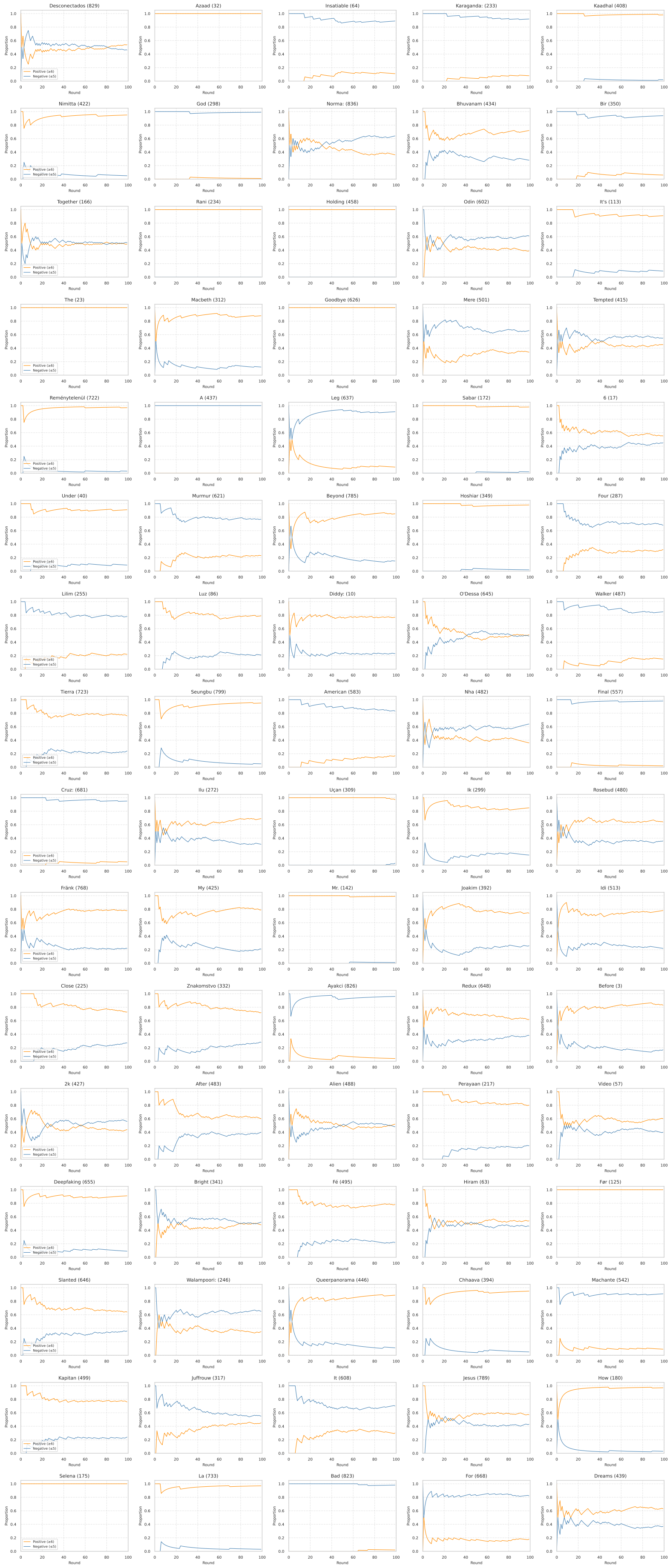


Figure 12: Trend of Positive and Negative Opinion Proportions on GPT-4o-mini under the “w/ Persona & w/o History” Setting.

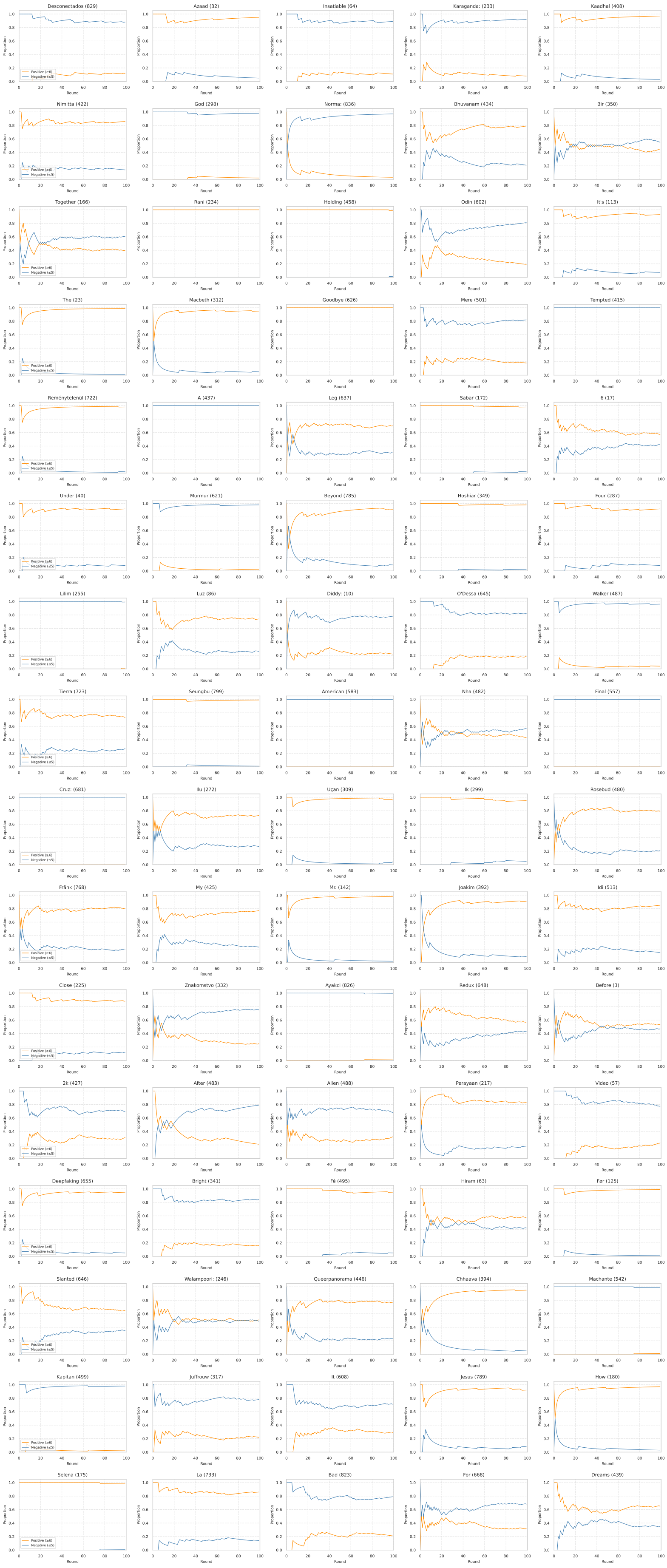


Figure 13: Trend of Positive and Negative Opinion Proportions on GPT-4o-mini under the “w/ Persona & w/ History” Setting.

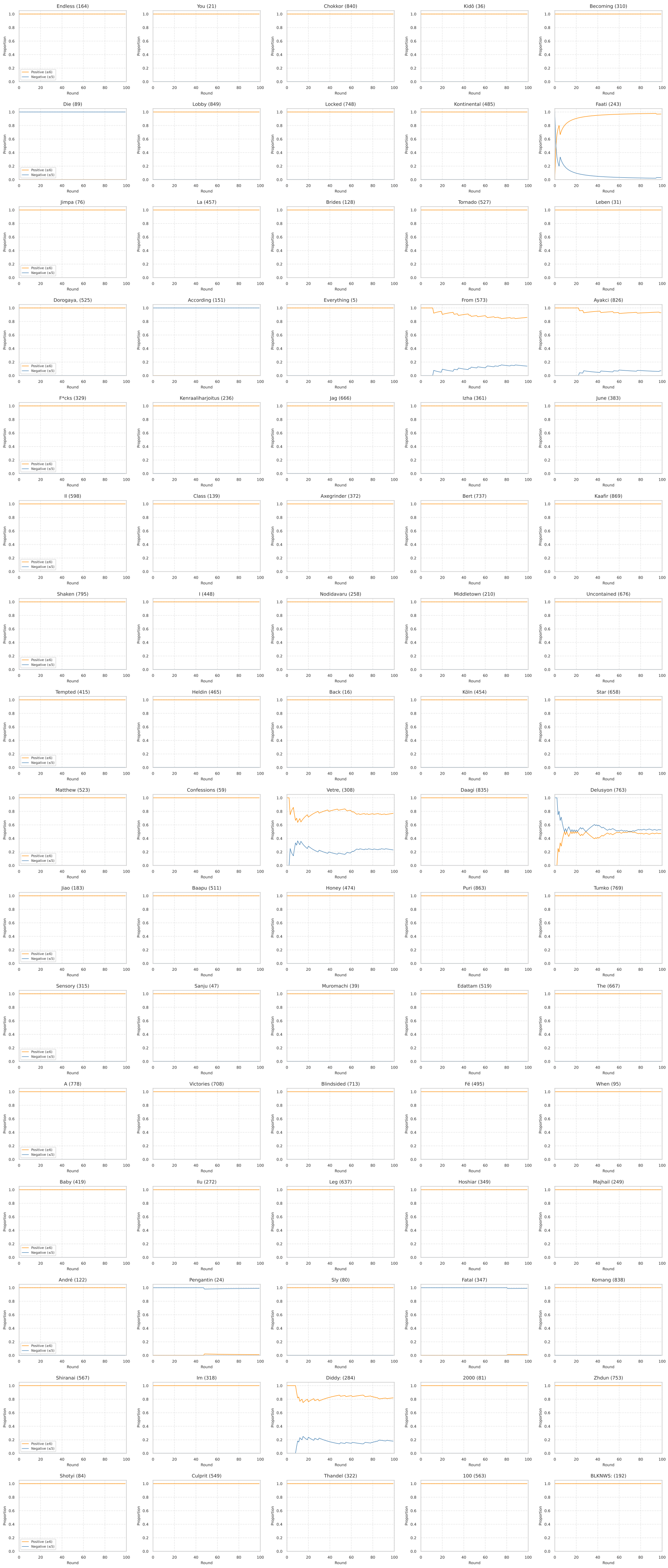


Figure 14: Trend of Positive and Negative Opinion Proportions on DeepSeek-V2-Lite-Chat under the “w/o Persona & w/o History” Setting.

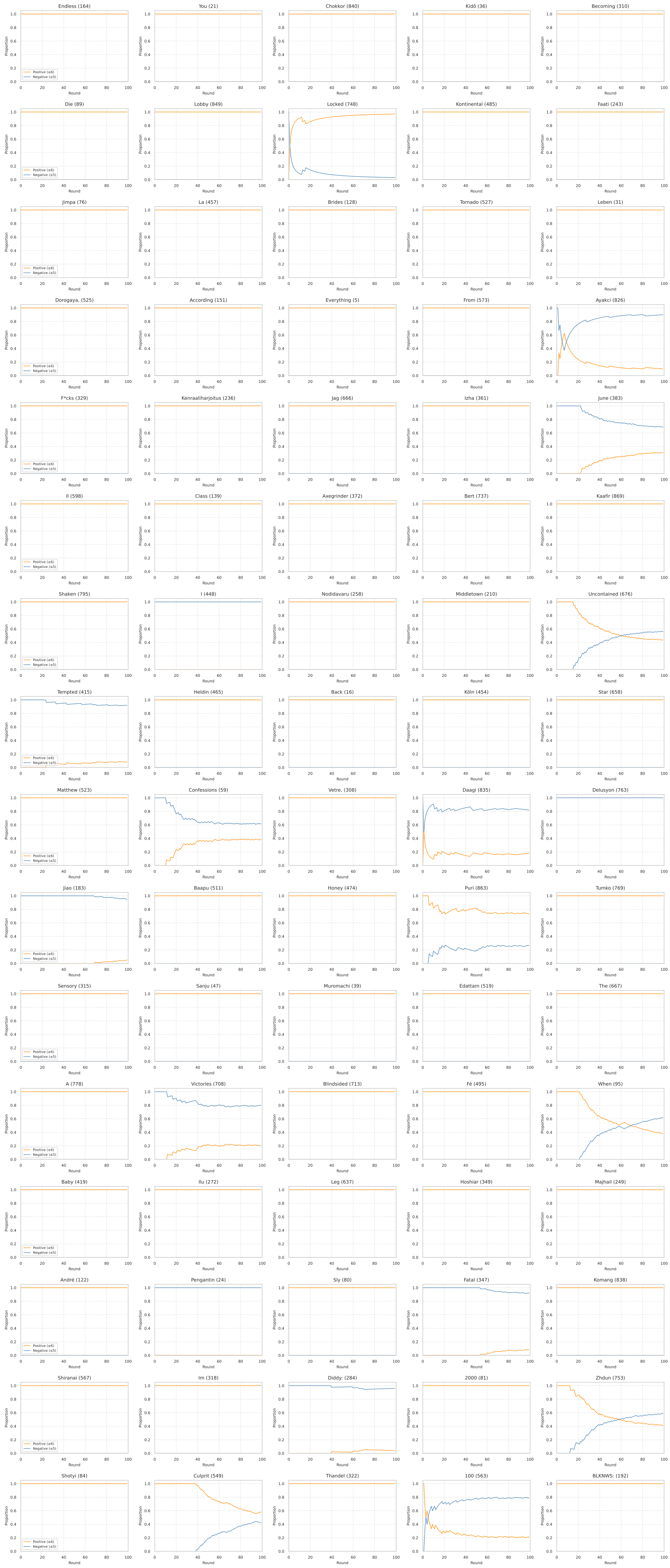
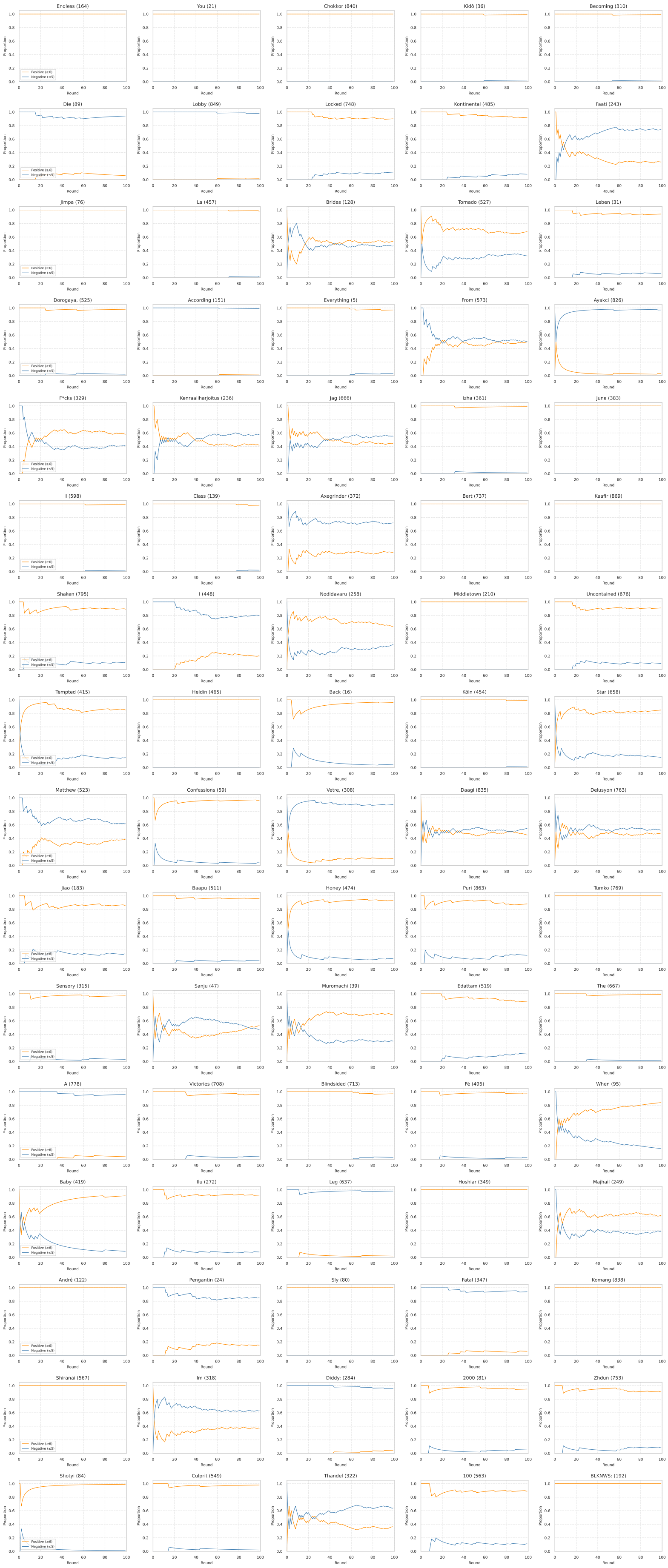


Figure 15: Trend of Positive and Negative Opinion Proportions on DeepSeek-V2-Lite-Chat under the "w/o Persona & w/ History" Setting.



Endless (164)
You (21)
Chokkor (840)
Kidô (36)
Becoming (310)
Die (89)
Lobby (849)
Locked (748)
Kontinental (485)
Faati (243)
Jimpa (76)
La (457)
Brides (128)
Tornado (527)
Leben (31)
Dorogaya, (525)
According (151)
Everything (5)
From (573)
Ayakci (826)
F*cks (329)
Kenraaliharjoitus (236)
Jag (666)
Izha (361)
June (383)
Il (598)
Class (139)
Axegrinder (372)
Bert (737)
Kaafir (869)
Shaken (795)
I (448)
Nodidavaru (258)
Middletown (210)
Uncontained (676)
Tempted (415)
Heldin (465)
Back (16)
Köln (454)
Star (658)
Matthew (523)
Confessions (59)
Vetre, (308)
Daagi (835)
Delusyon (763)
Jiao (183)
Baapu (511)
Honey (474)
Puri (863)
Tumko (769)
Sensory (315)
Sanju (47)
Muromachi (39)
Edattam (519)
The (667)
A (778)
Victories (708)
Blindsided (713)
Fé (495)
When (95)
Baby (419)
Ilu (272)
Leg (637)
Hoshiar (349)
Majhail (249)
André (122)
Pengantin (24)
Sly (80)
Fatal (347)
Komang (838)
Shiranai (567)
Im (318)
Diddy: (284)
2000 (81)
Zhdun (753)
Shotyi (84)
Culprit (549)
Thandel (322)
100 (563)
BLKNWS: (192)

Figure 16: Trend of Positive and Negative Opinion Proportions on DeepSeek-V2-Lite-Chat under the “w/ Persona & w/o History” Setting.

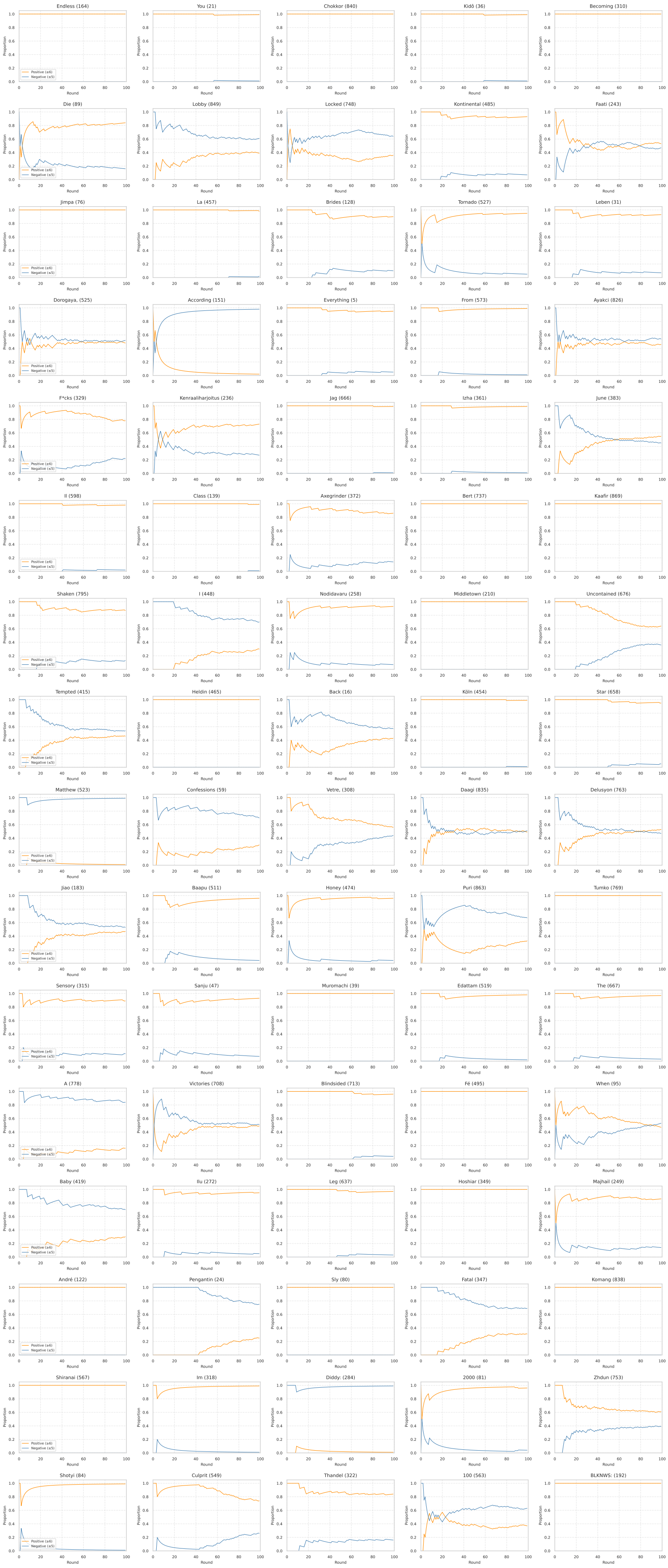


Figure 17: Trend of Positive and Negative Opinion Proportions on DeepSeek-V2-Lite-Chat under the “w/ Persona & w/ History” Setting.

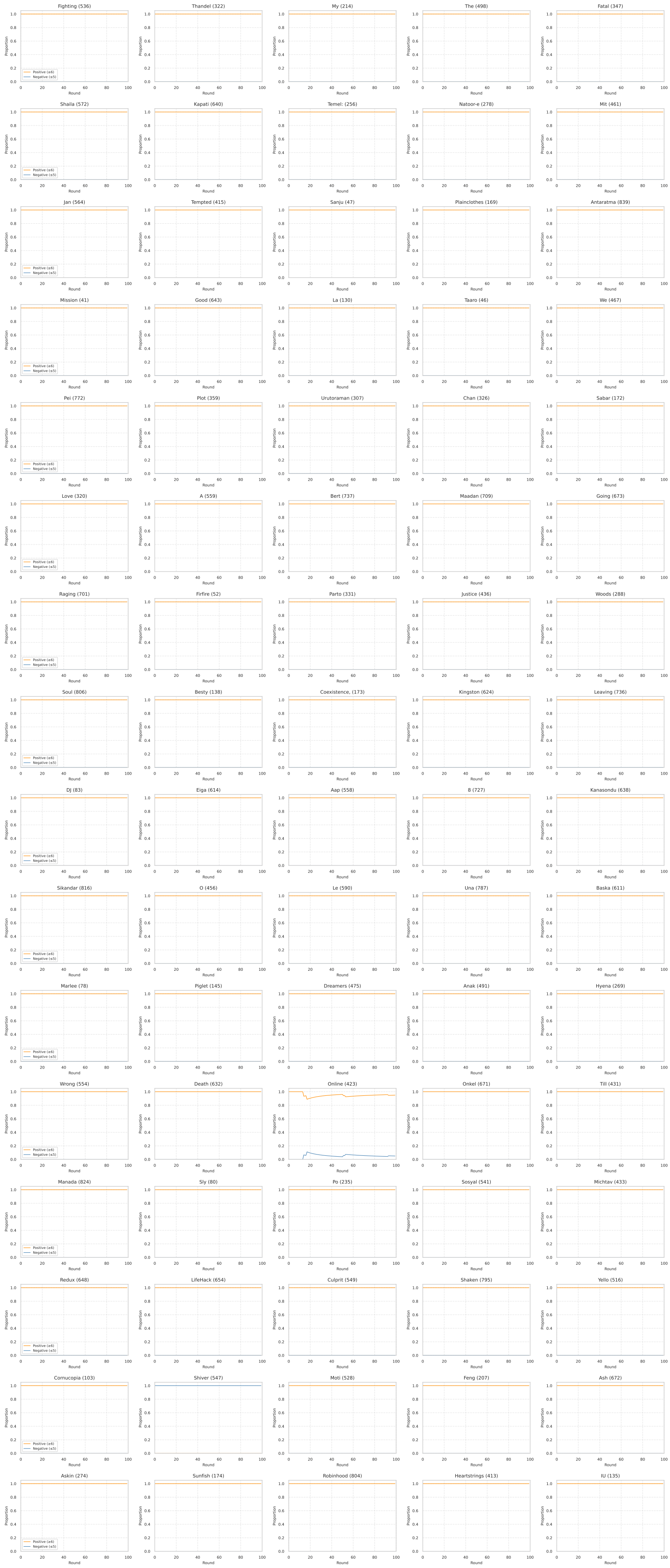


Figure 18: Trend of Positive and Negative Opinion Proportions on Ministral-8B-Instruct-2410 under the “w/o Persona & w/o History” Setting.

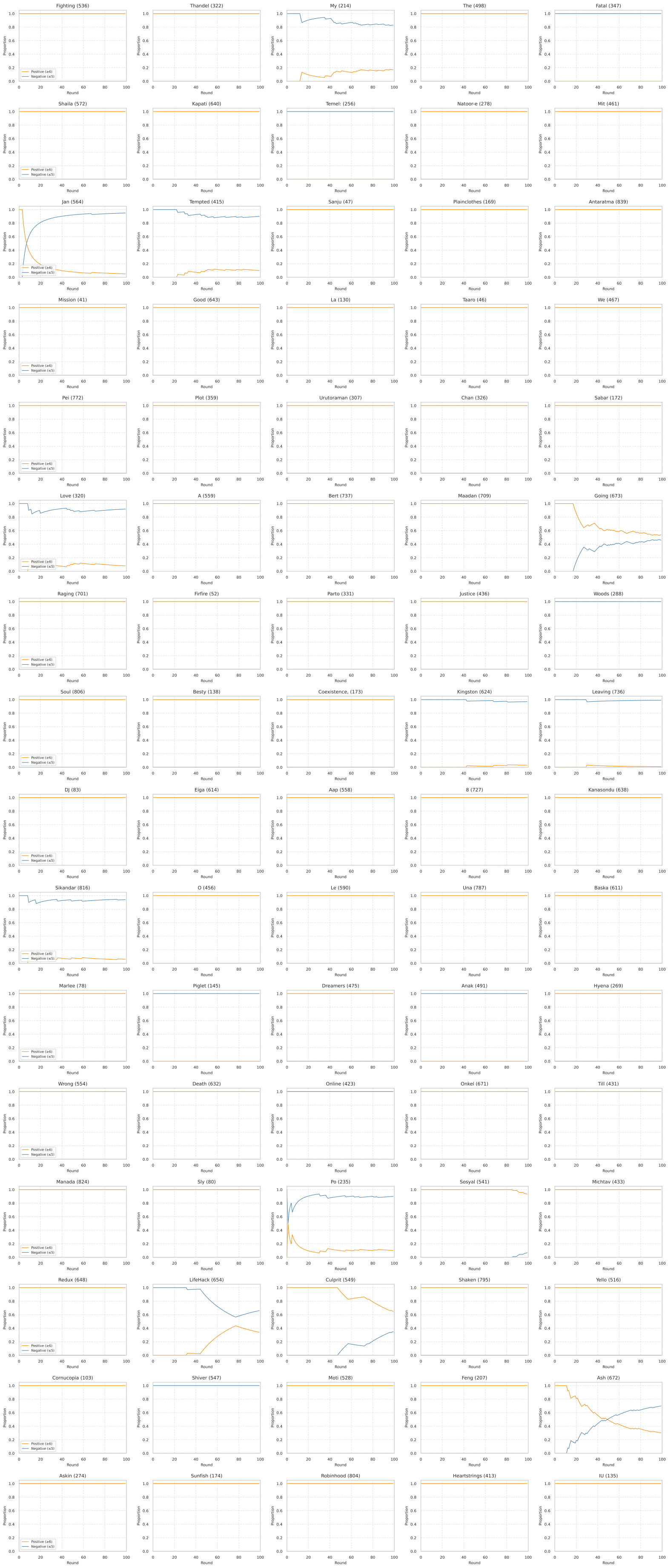


Figure 19: Trend of Positive and Negative Opinion Proportions on Ministral-8B-Instruct-2410 under the "w/o Persona & w/ History" Setting.

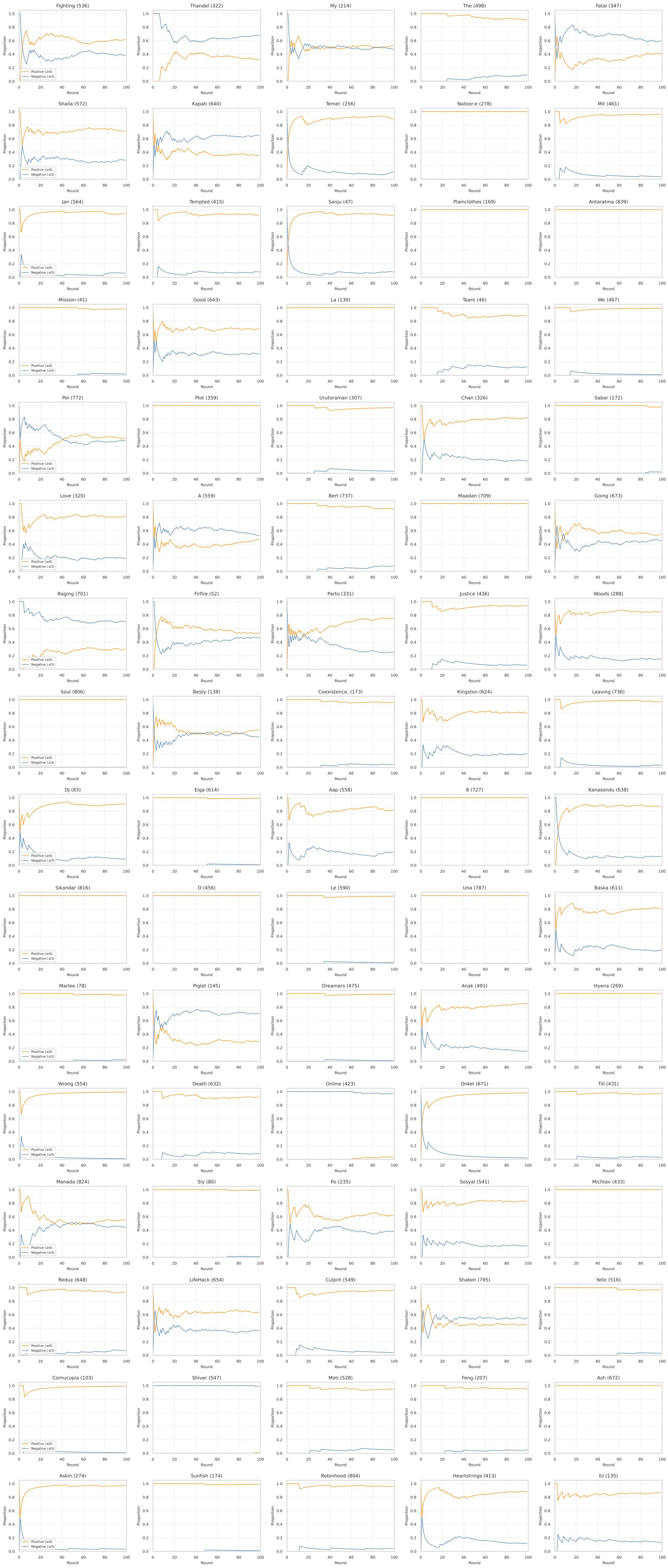


Figure 20: Trend of Positive and Negative Opinion Proportions on Ministral-8B-Instruct-2410 under the “w/ Persona & w/o History” Setting.

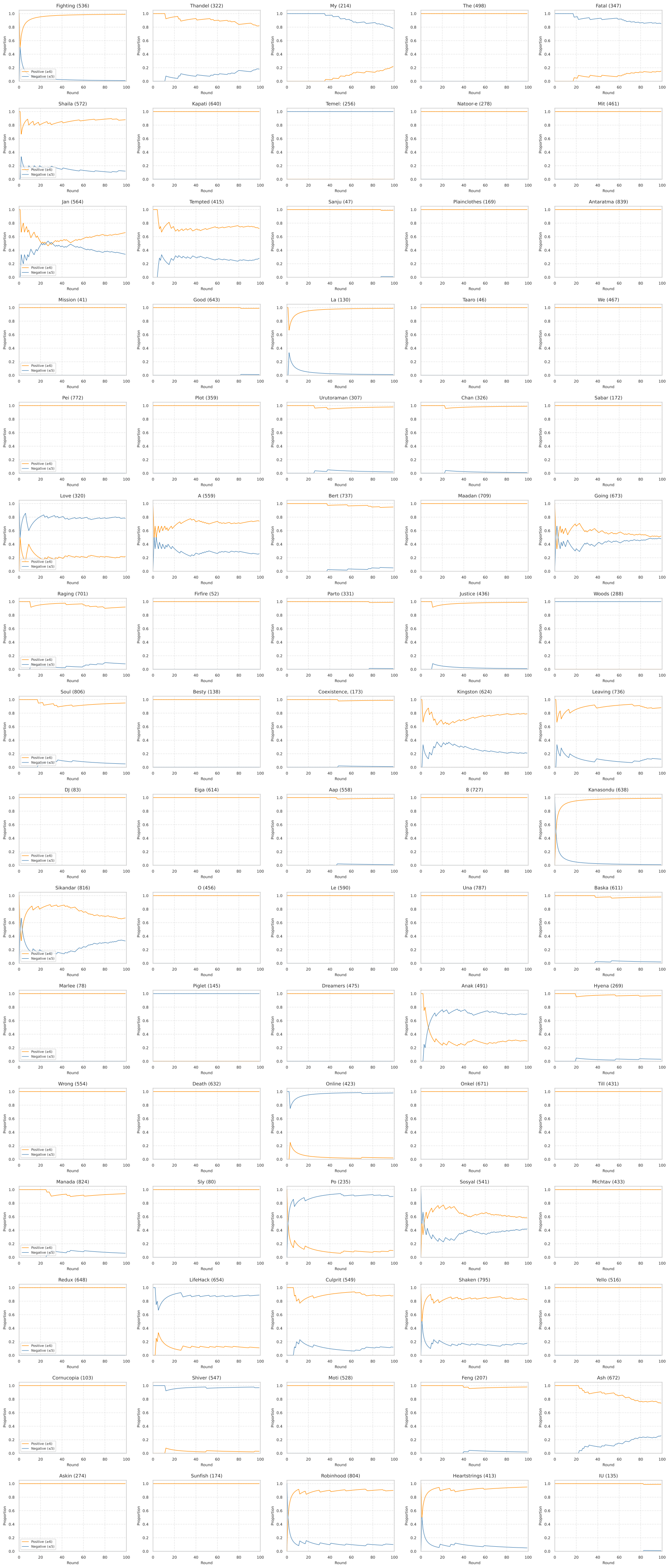


Figure 21: Trend of Positive and Negative Opinion Proportions on Ministral-8B-Instruct-2410 under the "w/ Persona & w/ History" Setting.

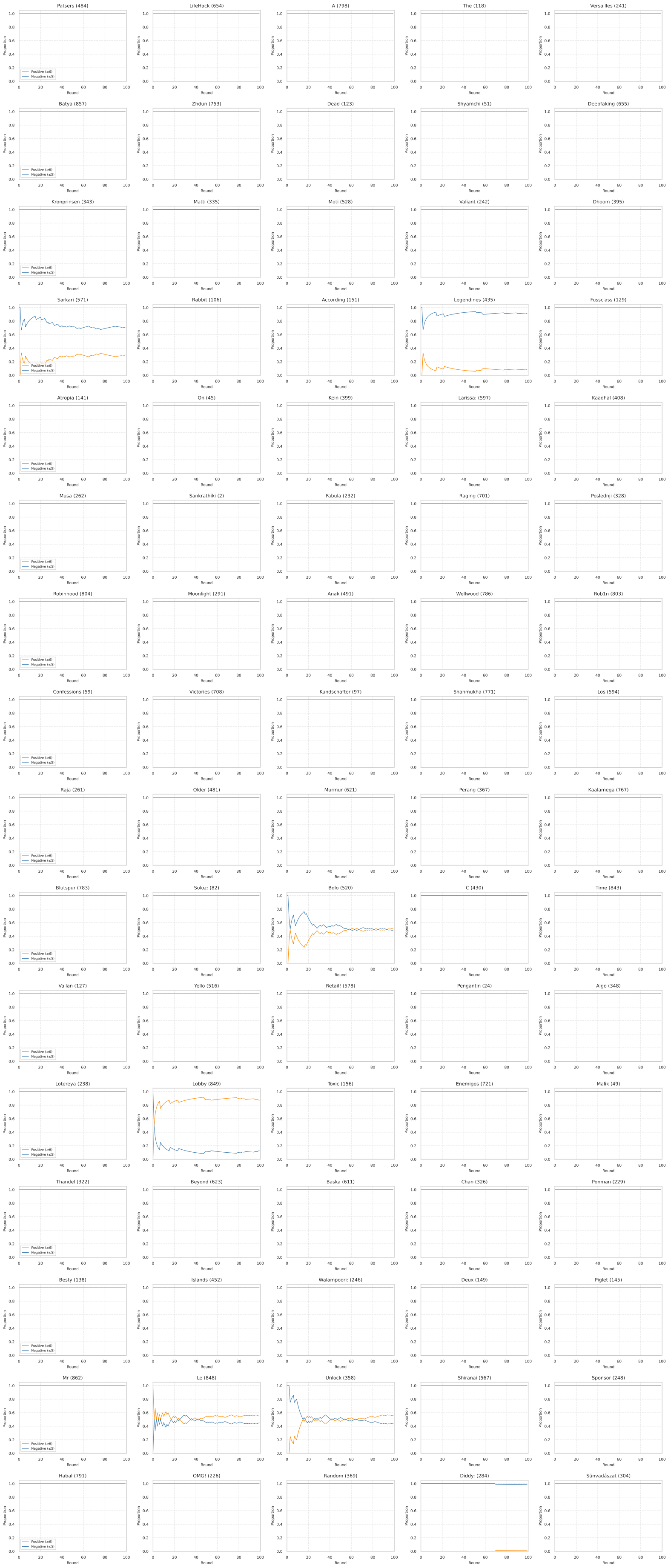


Figure 22: Trend of Positive and Negative Opinion Proportions on Qwen2.5-1.5B-Instruct under the "w/o Persona & w/o History" Setting.

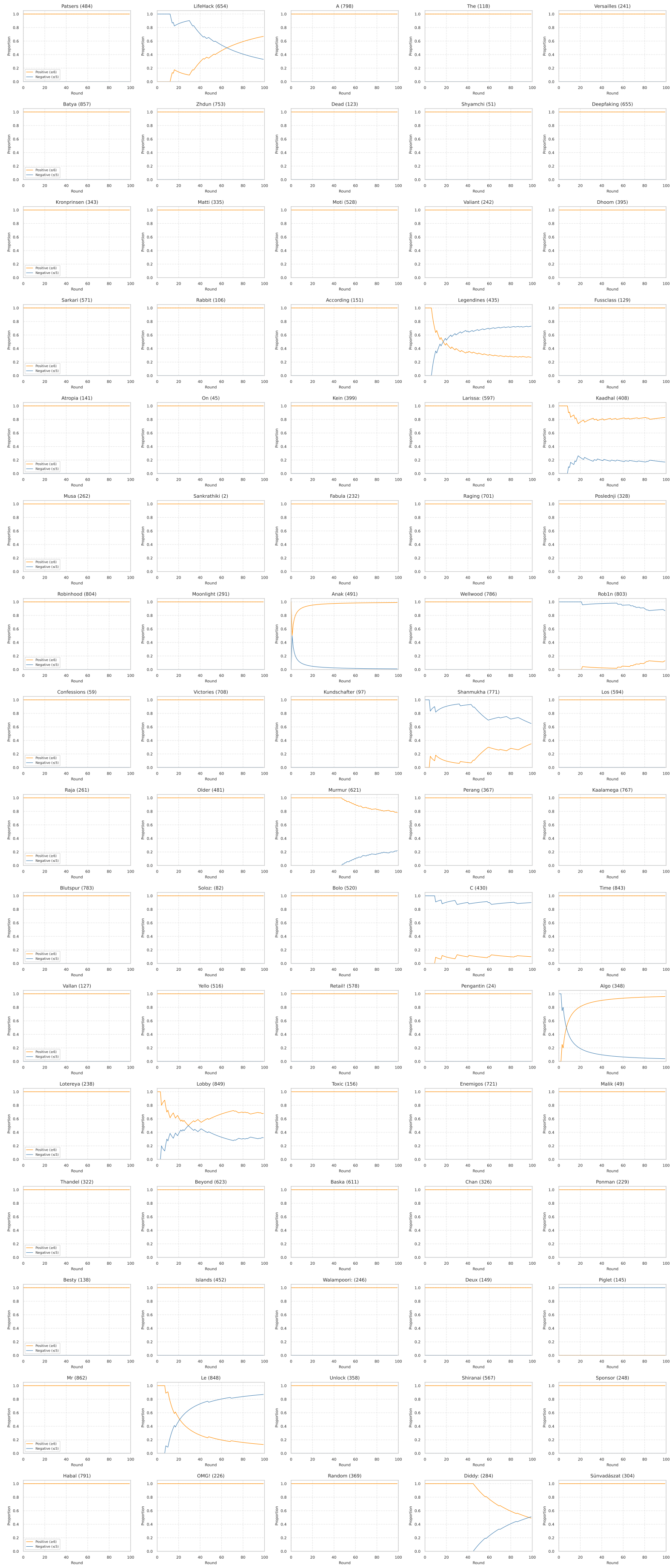


Figure 23: Trend of Positive and Negative Opinion Proportions on Qwen2.5-1.5B-Instruct under the “w/o Persona & w/ History” Setting.

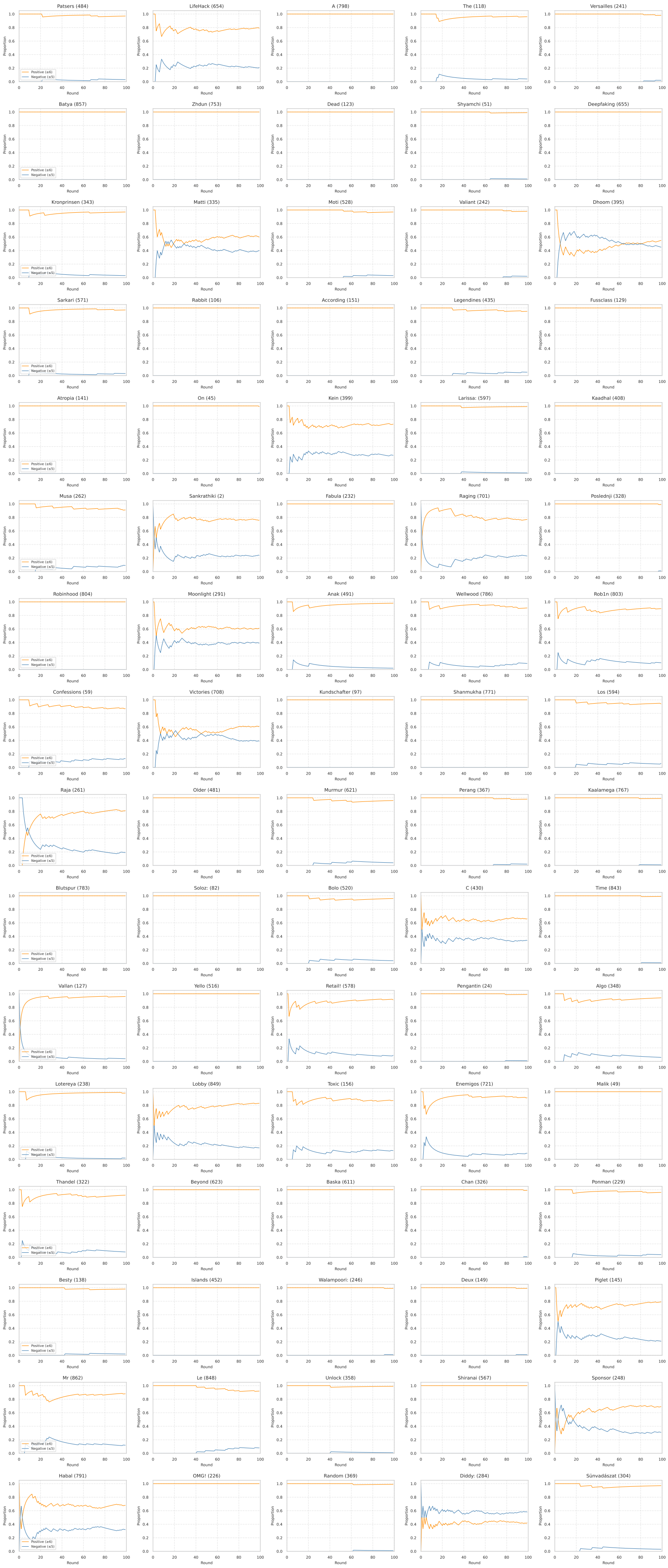


Figure 24: Trend of Positive and Negative Opinion Proportions on Qwen2.5-1.5B-Instruct under the “w/ Persona & w/o History” Setting.

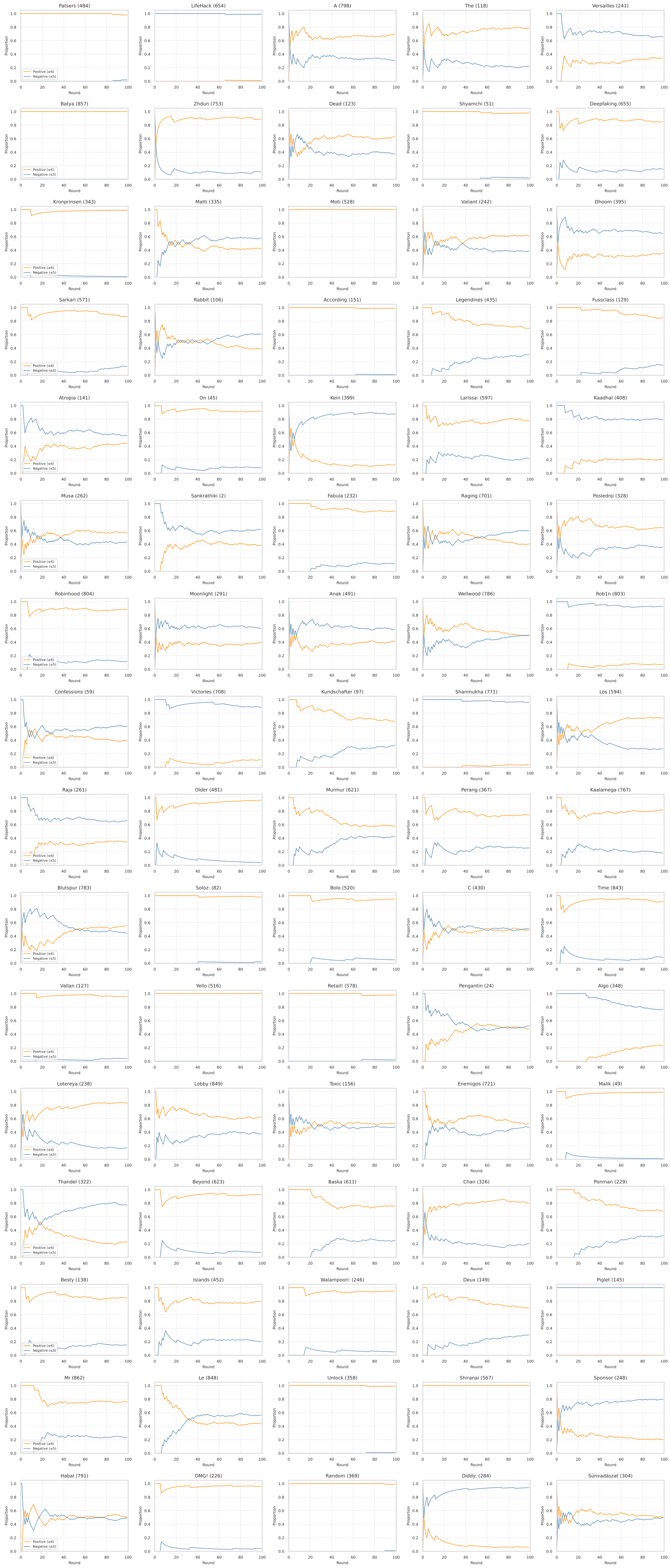


Figure 25: Trend of Positive and Negative Opinion Proportions on Qwen2.5-1.5B-Instruct under the “w/ Persona & w/ History” Setting.

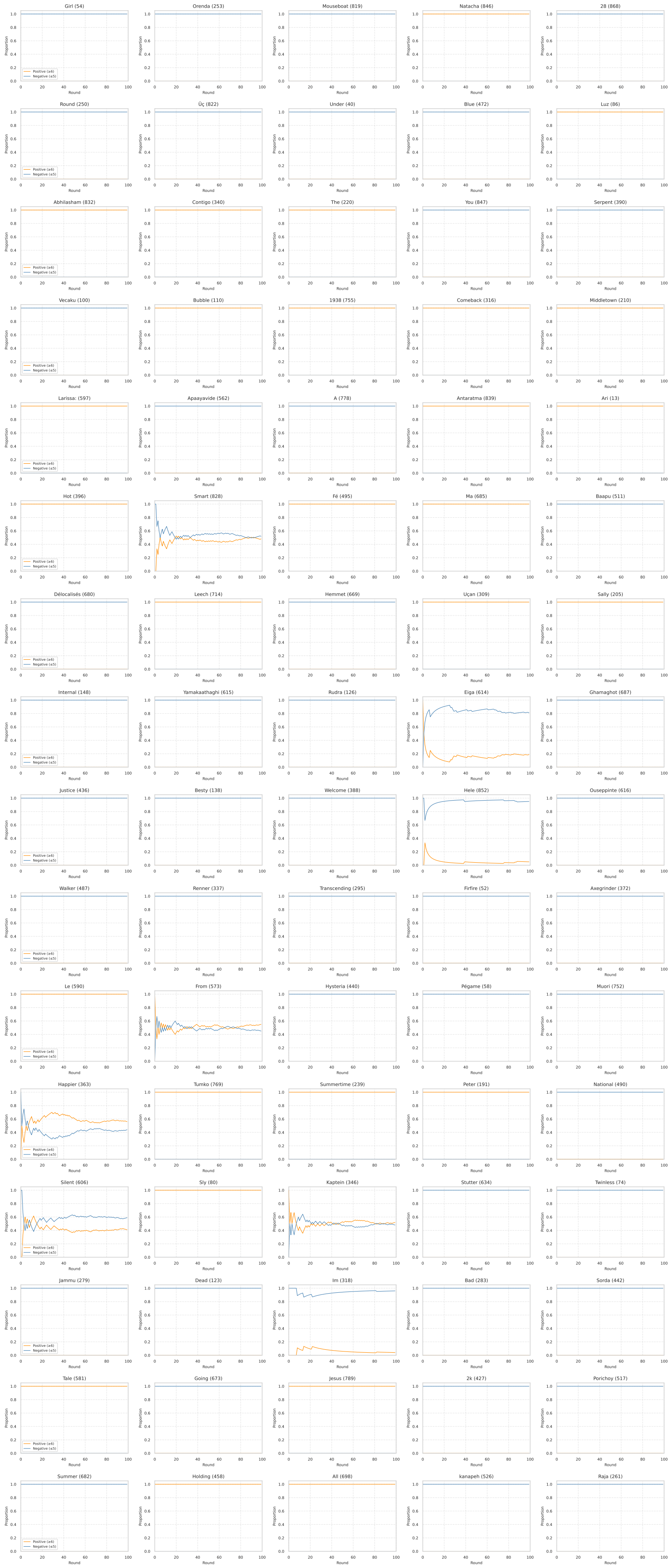


Figure 26: Trend of Positive and Negative Opinion Proportions on Qwen2.5-3B-Instruct under the “w/o Persona & w/o History” Setting.

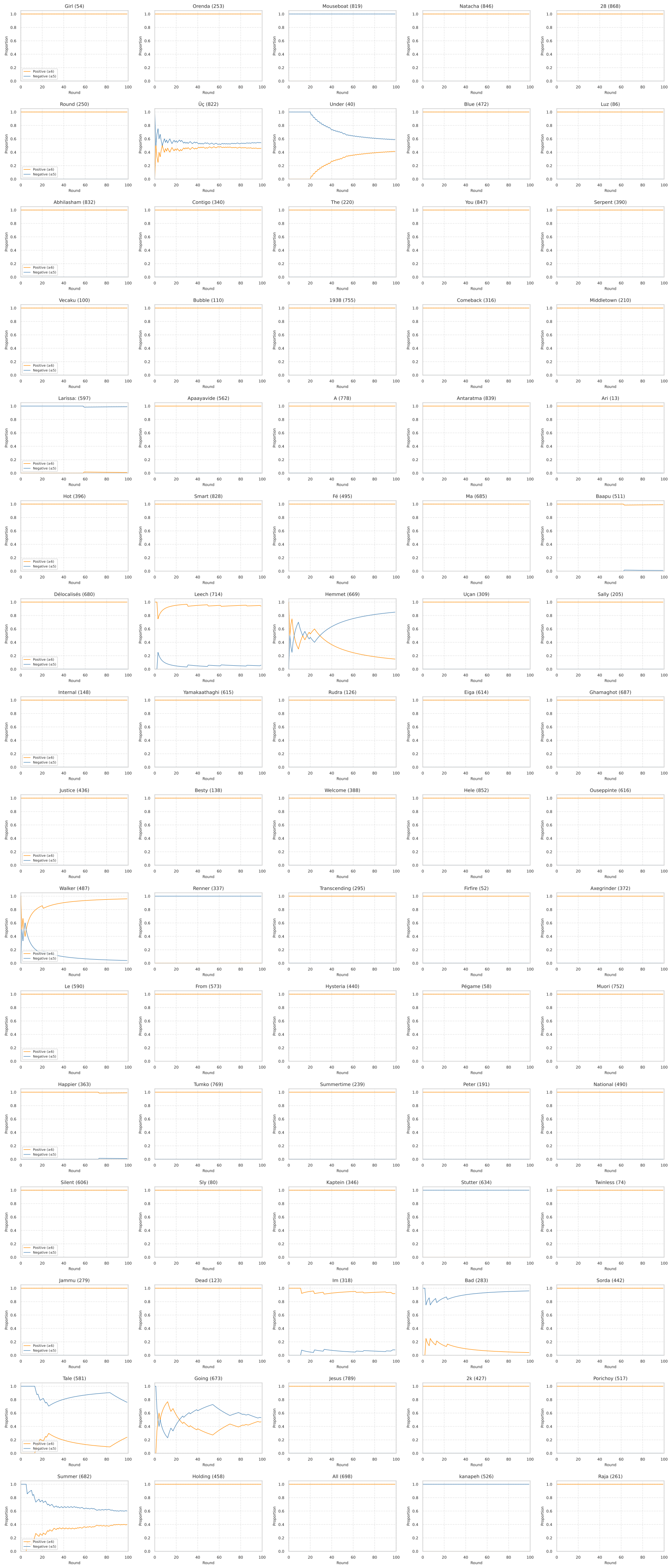


Figure 27: Trend of Positive and Negative Opinion Proportions on Qwen2.5-3B-Instruct under the “w/o Persona & w/ History” Setting.

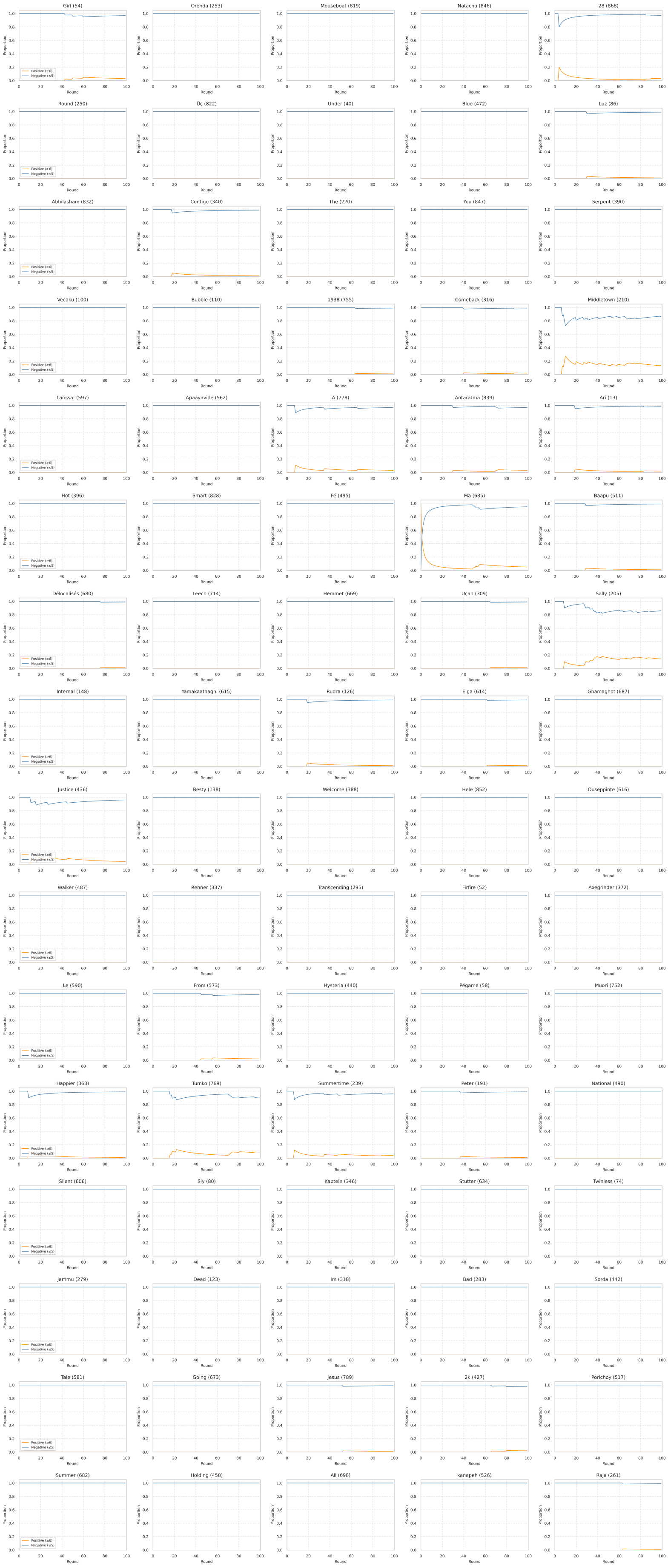


Figure 28: Trend of Positive and Negative Opinion Proportions on Qwen2.5-3B-Instruct under the “w/ Persona & w/o History” Setting.

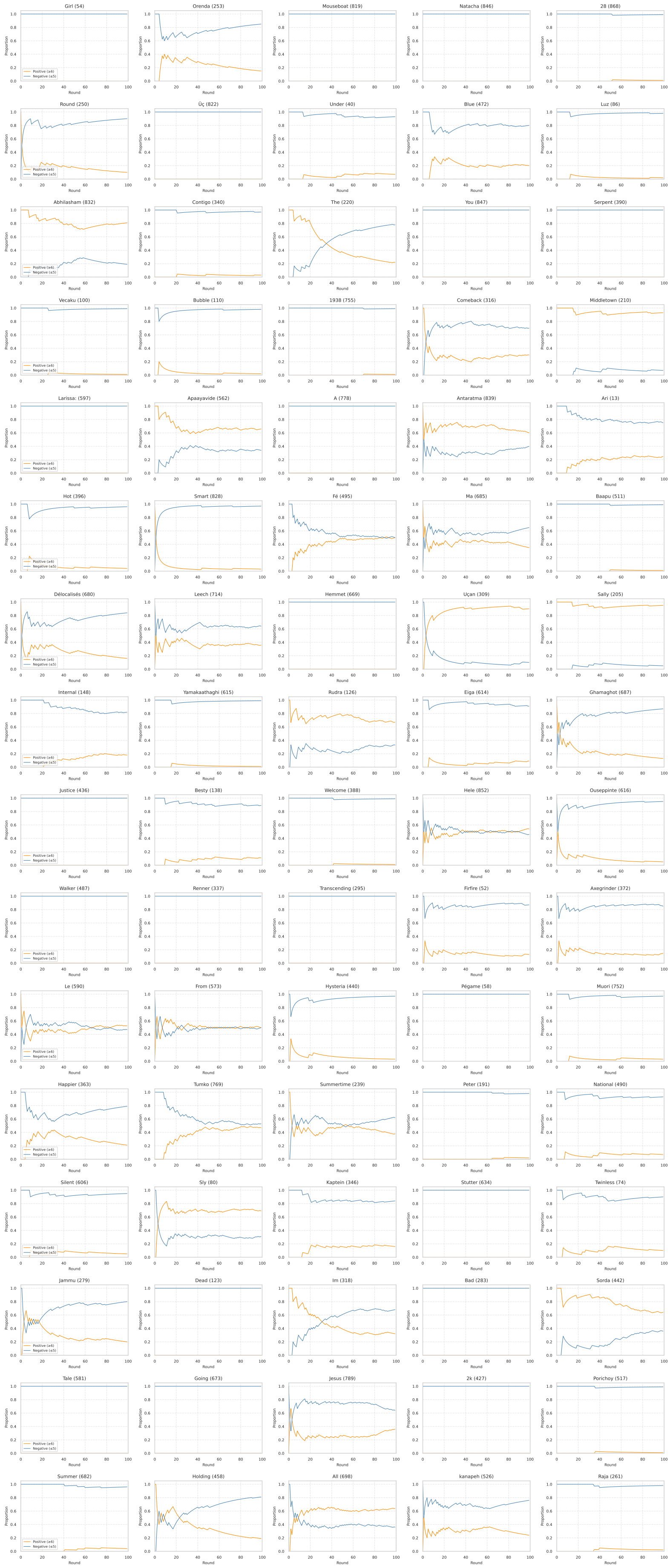


Figure 29: Trend of Positive and Negative Opinion Proportions on Qwen2.5-3B-Instruct under the "w/ Persona & w/ History" Setting.

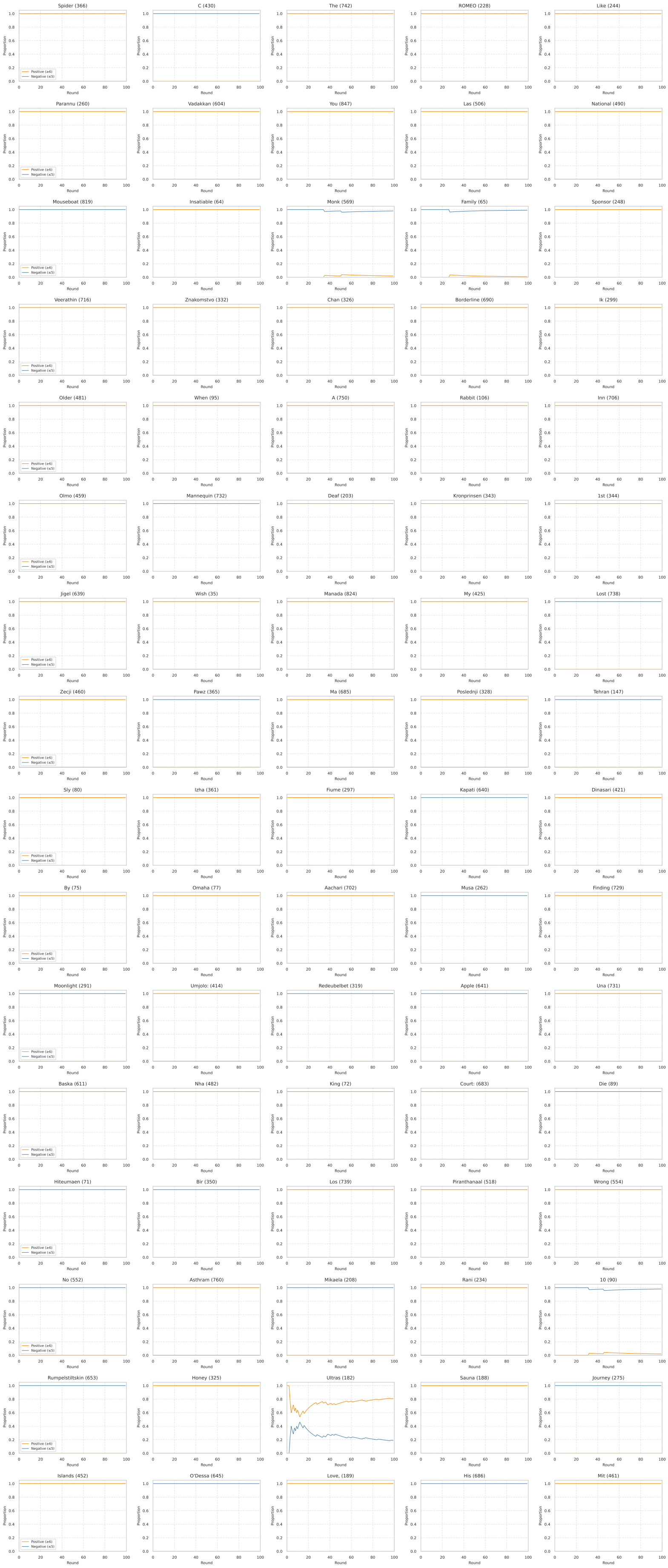
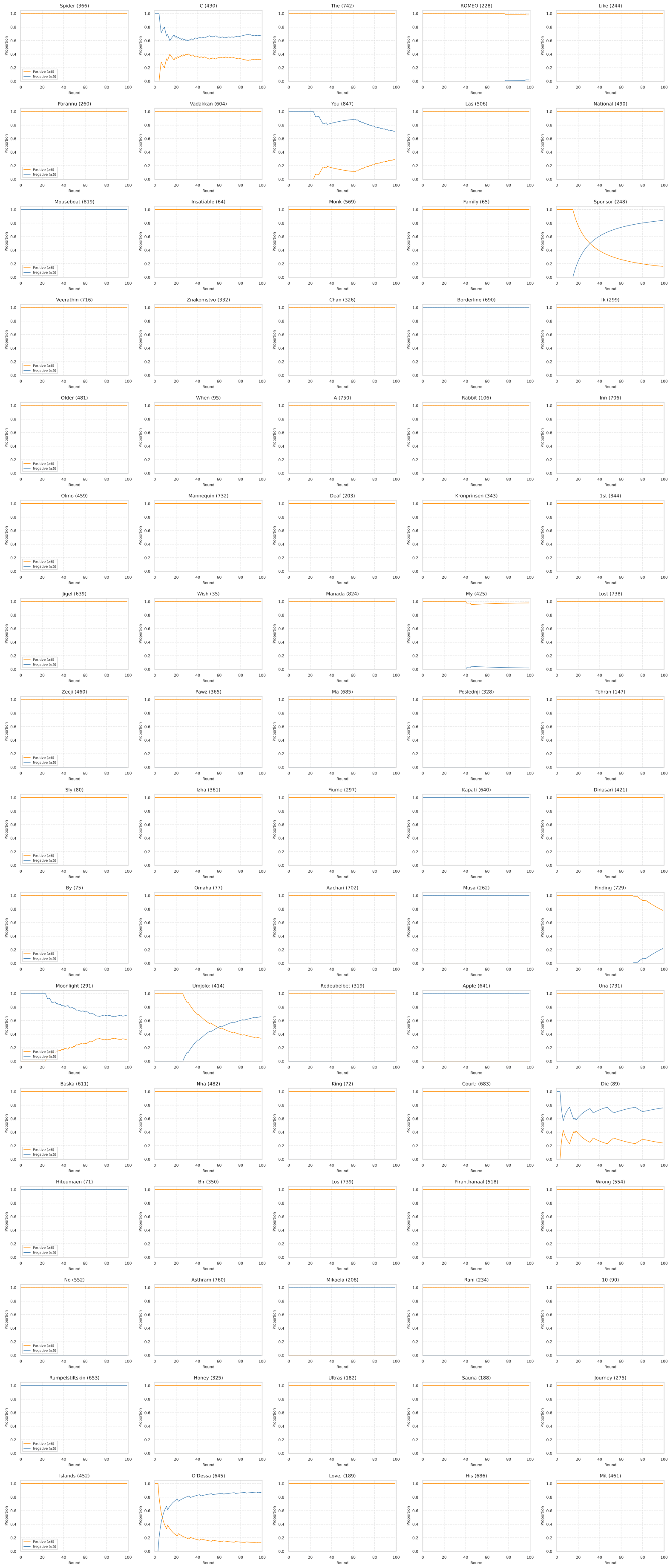


Figure 30: Trend of Positive and Negative Opinion Proportions on Qwen2.5-7B-Instruct under the “w/o Persona & w/o History” Setting.



Spider (366) C (430) The (742) ROMEO (228) Like (244)

Parannu (260) Vadakkan (604) You (847) Las (506) National (490)

Mouseboat (819) Insatiable (64) Monk (569) Family (65) Sponsor (248)

Veerathin (716) Znakomstvo (332) Chan (326) Borderline (690) Ik (299)

Older (481) When (95) A (750) Rabbit (106) Inn (706)

Olmo (459) Mannequin (732) Deaf (203) Kronprinsen (343) 1st (344)

Jigel (639) Wish (35) Manada (824) My (425) Lost (738)

Zecji (460) Pawz (365) Ma (685) Poslednji (328) Tehran (147)

Sly (80) Izha (361) Fiume (297) Kapati (640) Dinasari (421)

By (75) Omaha (77) Aachari (702) Musa (262) Finding (729)

Moonlight (291) Umjolo: (414) Redeubelbet (319) Apple (641) Una (731)

Baska (611) Nha (482) King (72) Court: (683) Die (89)

Hiteumaen (71) Bir (350) Los (739) Piranthanaal (518) Wrong (554)

No (552) Asthram (760) Mikaela (208) Rani (234) 10 (90)

Rumpelstiltskin (653) Honey (325) Ultras (182) Sauna (188) Journey (275)

Islands (452) O'Dessa (645) Love, (189) His (686) Mit (461)

Figure 31: Trend of Positive and Negative Opinion Proportions on Qwen2.5-7B-Instruct under the “w/o Persona & w/ History” Setting.

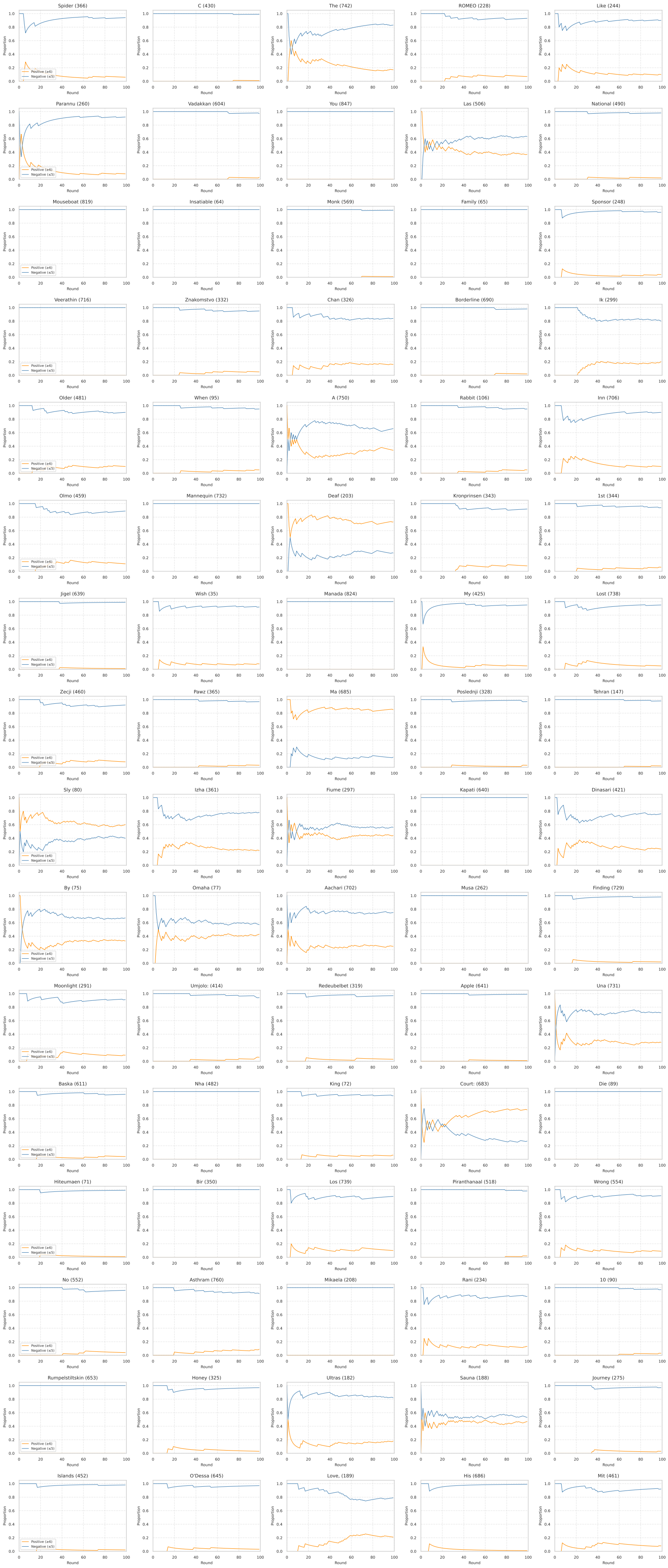


Figure 32: Trend of Positive and Negative Opinion Proportions on Qwen2.5-7B-Instruct under the “w/ Persona & w/o History” Setting.

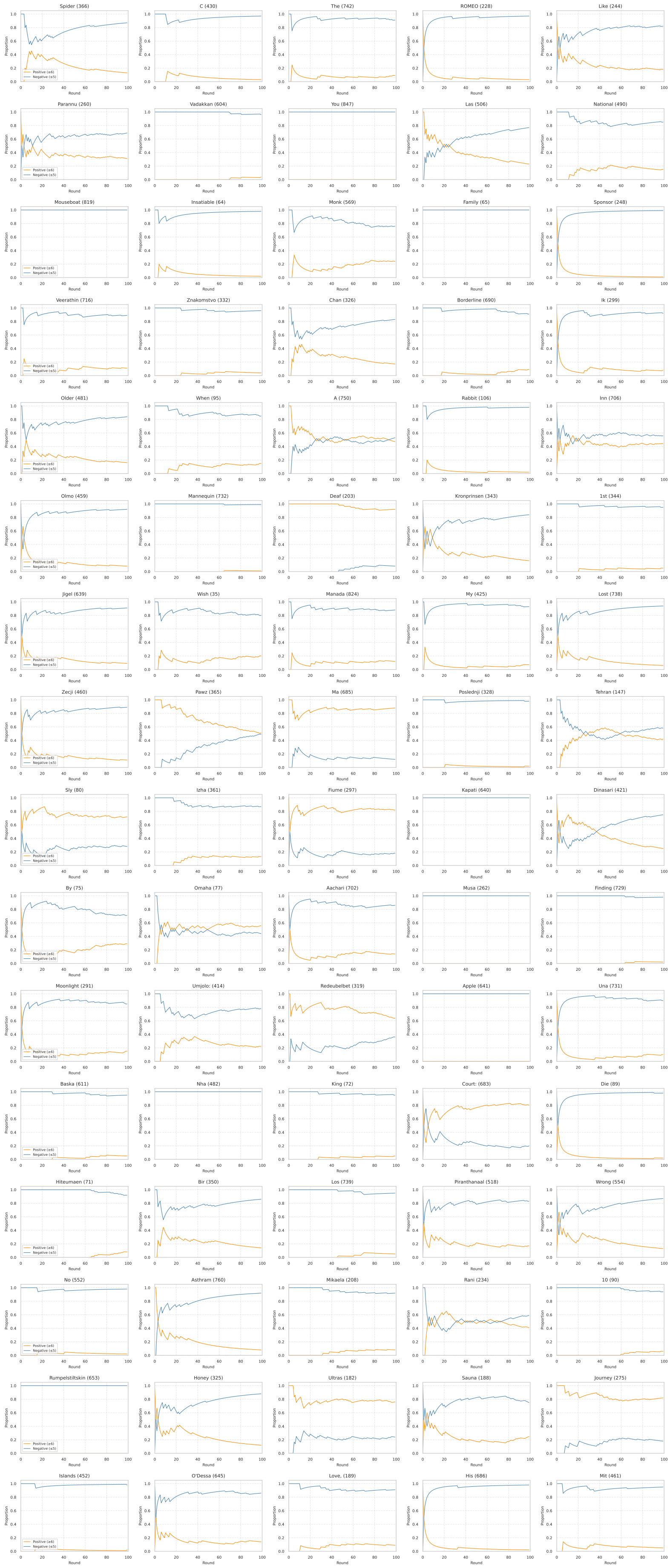


Figure 33: Trend of Positive and Negative Opinion Proportions on Qwen2.5-7B-Instruct under the “w/ Persona & w/ History” Setting.

A.5 Semantic Similarity vs. Rating Distance (w/ Persona & w/ History)

In the “w/ Persona & w/ History” setting, we examined scatter plots of TF-IDF semantic similarity vs. rating distance. With the x-axis showing the TF-IDF similarity between the agent’s persona description and the movie overview, and the y-axis showing the rating distance (absolute deviation of the agent’s rating from its historical average rating for movies). Across all evaluated language models (except LLaMA-3.1-8B-Instruct), we observe a consistent inverse relationship between persona–movie similarity and rating distance. Specifically, lower persona–movie similarity scores correspond to higher rating distances (that is, when an agent’s persona is less aligned with a movie, the agent’s rating tends to deviate more from its usual average). In contrast, movies that are more similar to the persona of the agent yield ratings closer to the historical average of the agent (smaller deviation on the y axis). This same pattern is evident for every model tested – including DeepSeek-V2-Lite-Chat, Mistral-8B-Instruct-2410, and the Qwen-2.5-Instruct series – mirroring the trend originally observed with GPT-4o-mini.

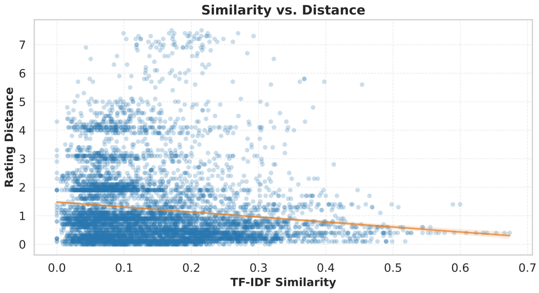


Figure 34: The relation between **Semantic Similarity vs. Rating Distance** on DeepSeek-V2-Lite-Chat.

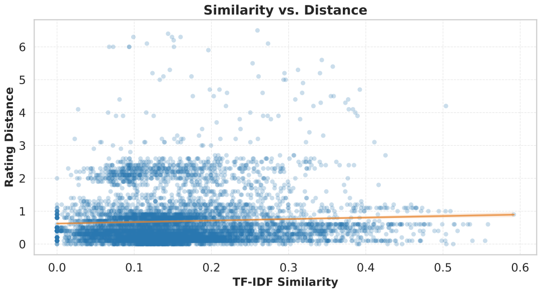


Figure 35: The relation between **Semantic Similarity vs. Rating Distance** on Ministral-8B-Instruct-2410.

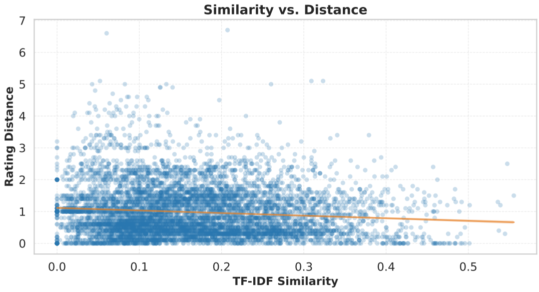


Figure 36: The relation between **Semantic Similarity vs. Rating Distance** on Qwen2.5-1.5B-Instruct.

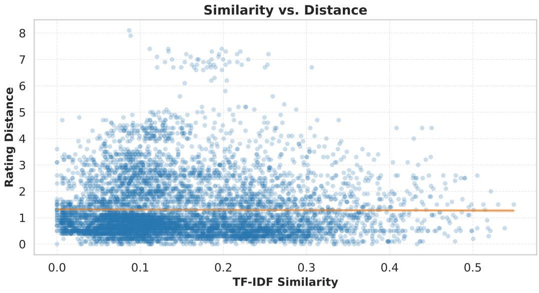


Figure 37: The relation between **Semantic Similarity vs. Rating Distance** on Qwen2.5-3B-Instruct.

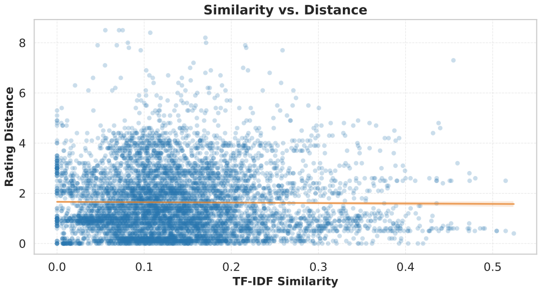
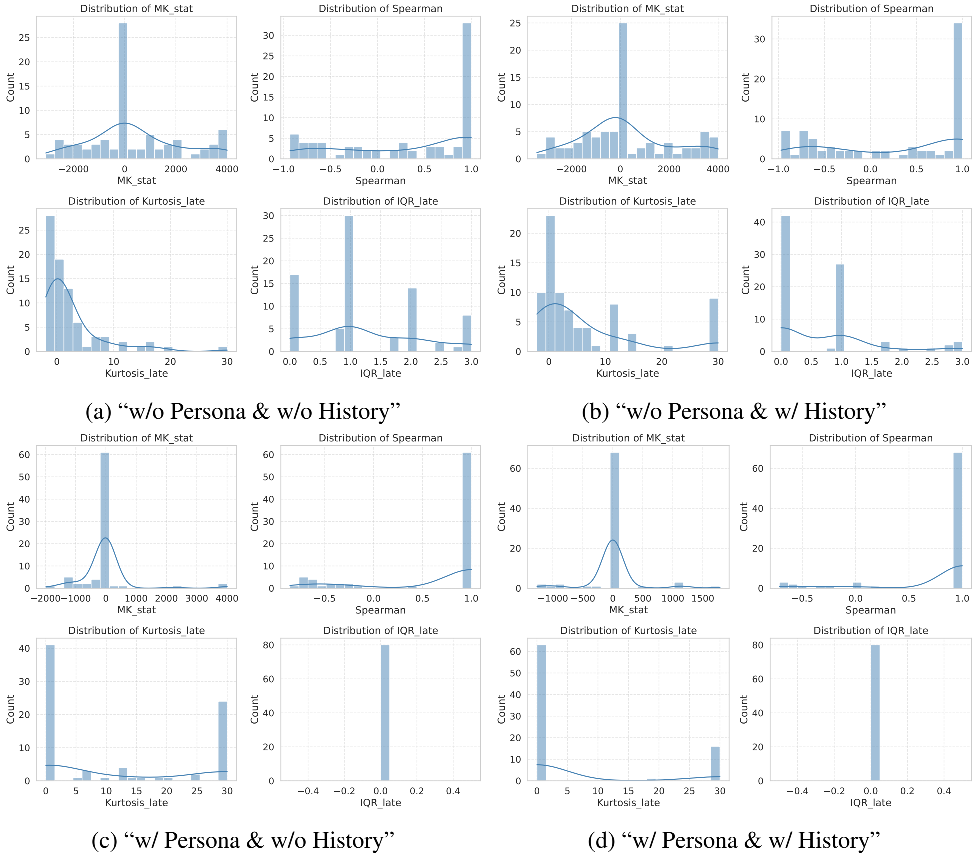


Figure 38: The relation between **Semantic Similarity vs. Rating Distance** on Qwen2.5-7B-Instruct.

A.6 Anomalous Case: Llama-3.1-8B-Instruct

This section summarizes the results of abnormal cases for model Llama-3.1-8B-Instruct. Llama-3.1-8B-Instruct exhibits a strikingly different pattern from the closed and most open-source models. In the with persona setting, it almost never forms a SoS: The MCO trend displays persistent fluctuations or coexistence of positive and negative views, with statistical indicators remaining dispersed and rarely showing the monotonic trends or sharp late-stage concentration found in other models. Even more surprisingly, in the “No Persona” setting , LLaMA-3.1-8B often shows strong fluctuations and sudden switches between positive and negative opinions. This atypical behavior may be due to multiple factors: greater diversity or stochasticity in the sampling of the model, weaker persona conditioning, an insufficient anchoring effect of persona / history inputs. Detailed results of the LLaMA-3.1-8B model are

A.6.1 Metrics Distributions



(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 39: Distributions of Mann–Kendall Statistic, Spearman Rank Correlation, Kurtosis, Inter-quartile Range for All Movie Rating Sequences on Llama-3.1-8B-Instruct.

A.6.2 Positive and Negative Opinion Proportion Trends (MCO Trend)

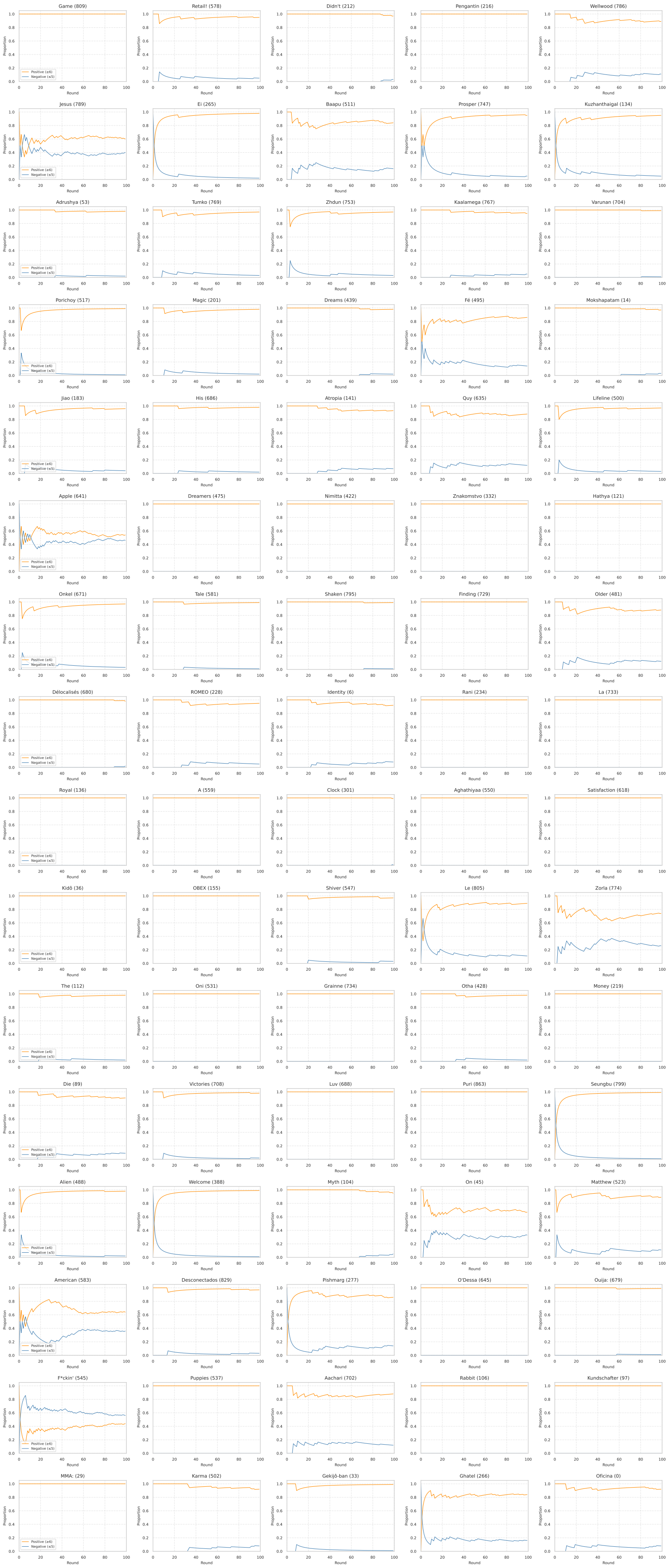


Figure 40: Trend of Positive and Negative Opinion Proportions on Llama-3.1-8B-Instruct under the “w/o Persona & w/o History” Setting.

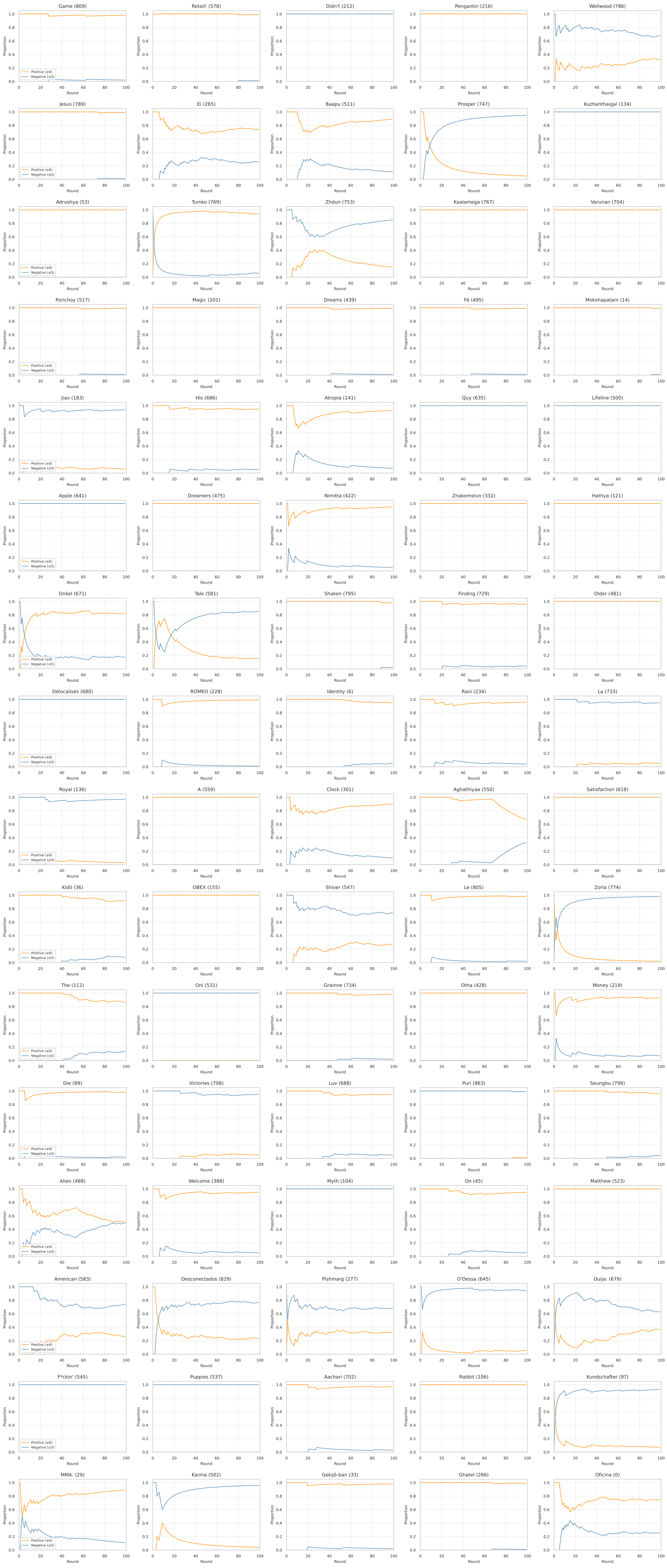


Figure 41: Trend of Positive and Negative Opinion Proportions on Llama-3.1-8B-Instruct under the “w/o Persona & w/ History” Setting.

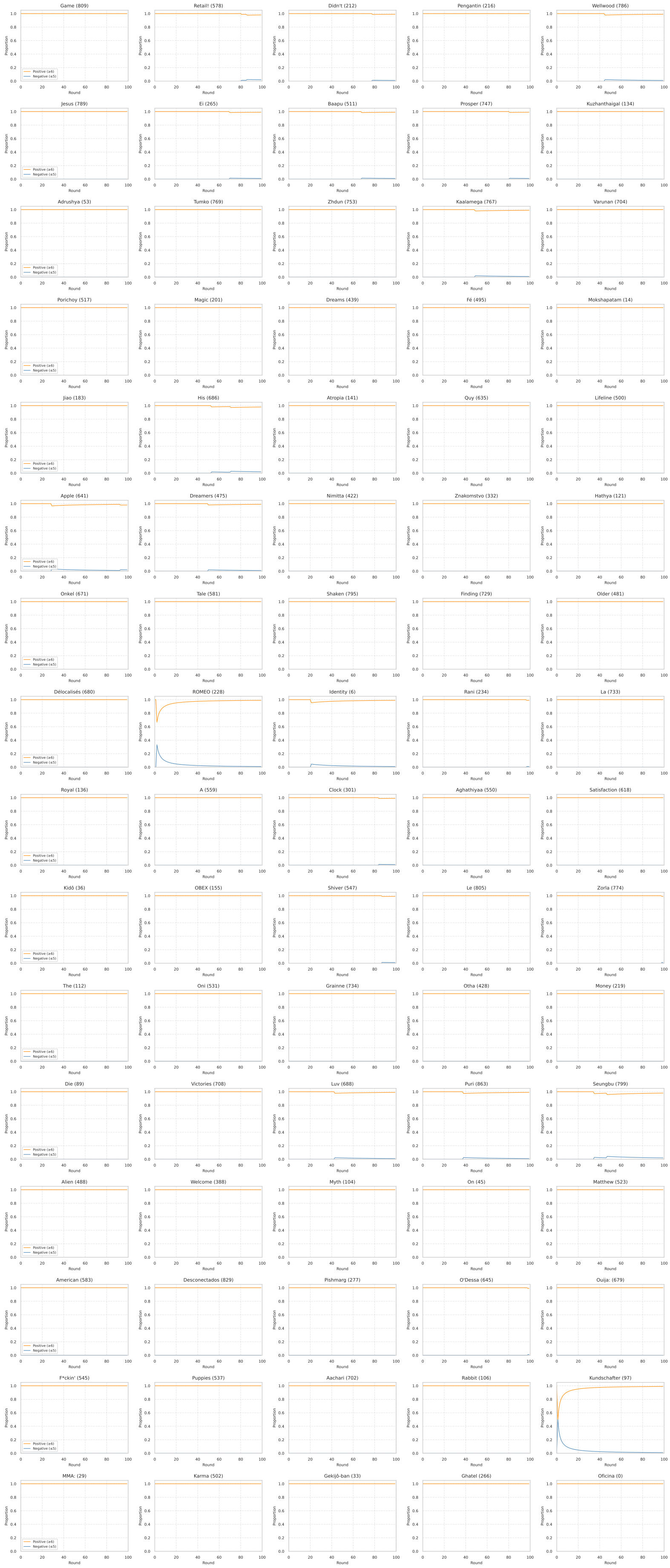


Figure 42: Trend of Positive and Negative Opinion Proportions on Llama-3.1-8B-Instruct under the "w/ Persona & w/o History" Setting.

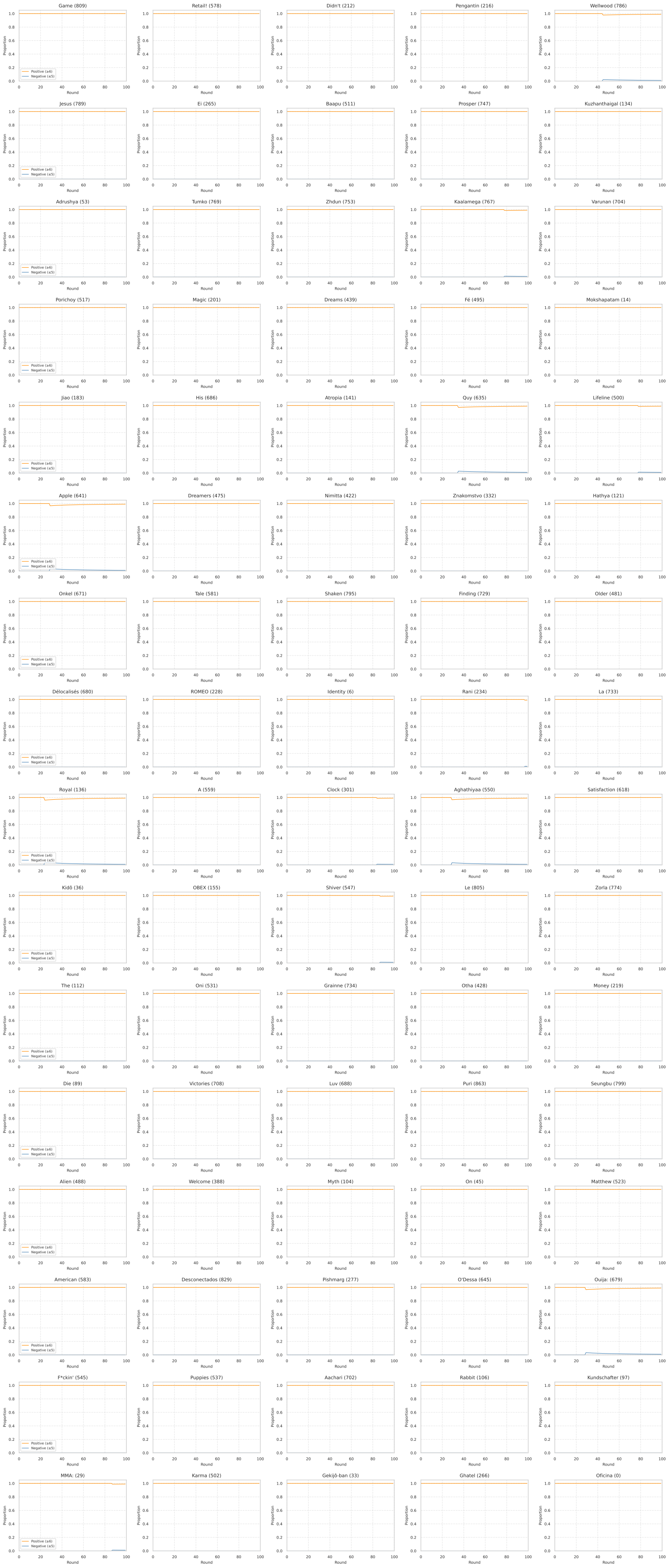


Figure 43: Trend of Positive and Negative Opinion Proportions on Llama-3.1-8B-Instruct under the "w/ Persona & w/ History" Setting.

A.6.3 Semantic Similarity vs. Rating Distance (w/ Persona & w/ History)

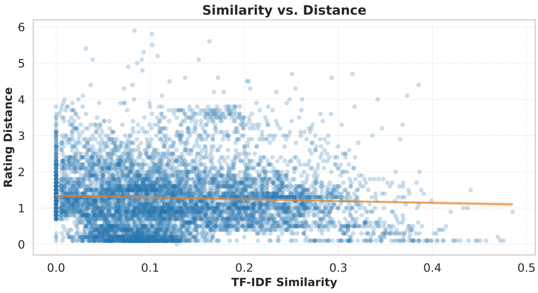


Figure 44: The relation between **Semantic Similarity vs. Rating Distance** on Llama-3.1-8B-Instruct.