Generative Exaggeration in LLM Social Agents: Consistency, Bias, and Toxicity

Jacopo Nudo[1*], Mario Edoardo Pandolfo[2†], Edoardo Loru[2†], Mattia Samory[1], Matteo Cinelli[1], Walter Quattrociocchi[1*]

[1*]Department of Computer Science, Sapienza University of Rome, Viale Regina Elena 295, Rome, 00161, Italy.
[2]Department of Computer, Control and Management Engineering, Sapienza University of Rome, Via Ariosto 25, Rome, 00185, Italy.

*Corresponding author(s). E-mail(s): jacopo.nudo@uniroma1.it; walter.quattrociocchi@uniroma1.it;
Contributing authors: marioedoardo.pandolfo@uniroma1.it; edoardo.loru@uniroma1.it; mattia.samory@uniroma1.it; matteo.cinelli@uniroma1.it;
†These authors contributed equally to this work.

Abstract

We investigate how Large Language Models (LLMs) behave when simulating political discourse on social media. Leveraging 21 million interactions on X during the 2024 U.S. presidential election, we construct LLM agents based on 1,186 real users, prompting them to reply to politically salient tweets under controlled conditions. Agents are initialized either with minimal ideological cues (Zero Shot) or recent tweet history (Few Shot), allowing one-to-one comparisons with human replies. We evaluate three model families—Gemini, Mistral, and DeepSeek—across linguistic style, ideological consistency, and toxicity. We find that richer contextualization improves internal consistency but also amplifies polarization, stylized signals, and harmful language. We observe an emergent distortion that we call "generation exaggeration": a systematic amplification of salient traits beyond empirical baselines. Our analysis shows that LLMs do not emulate users, they reconstruct them. Their outputs, indeed, reflect internal optimization dynamics more than observed behavior, introducing structural biases that compromise their reliability as social proxies. This challenges their use in content moderation, deliberative simulations, and policy modeling.

Introduction

The generative capacity of Large Language Models (LLMs) is currently used in a variety of applications, including customer support, search engines, educational tools, and content moderation. In many cases, LLMs do not just operate in isolation but are embedded in systems that require interaction or behavioral consistency. This shift—commonly referred to as agentification—consists of employing LLMs as agents that autonomously execute actions in interactive settings [1–6]. Agentified LLMs may perform social roles, sustain dialogue, and influence information flows within dynamic environments [7, 8].

Unlike traditional agent-based models [9, 10], which explicitly codify rules and assumptions to describe how individuals make decisions and interact, LLM-based agents generate behavior based on patterns extracted from large-scale training data. As a result, they may inherit the statistical biases present in the data and be influenced by its training procedure, rather than by any theoretical structure. This could have important consequences for how such agents behave, and how their outputs should be interpreted and evaluated.

In the near future, the prevalence of agentified LLMs is expected to expand [11], with LLMs taking on delegated roles in contexts where behavioral plausibility and social coherence matter. To understand the implications of this trend, it is necessary to examine what types of behavior LLMs produce and which distortions they may introduce into the environments in which they operate.

Social media platforms are a crucial testing ground to assess this shift. Their architecture already favors amplification of emotionally charged and polarizing content, often distorting political discourse and reinforcing tribal identities [12–15]. Using LLMs to simulate users in settings where political discussion happens in real time raises practical questions. What happens when generative agents are tasked with simulating human users engaged in political debate? Do they faithfully preserve the behavioral patterns of the profiles they emulate, or introduce new, systematic distortions?

In this study, we investigate the use of LLMs to simulate real users involved in the 2024 U.S. presidential election on X (formerly Twitter). We implement over 1,000 synthetic agents using six language models from three different model families—Gemini, Mistral, and DeepSeek—evaluating both larger and smaller variants to examine size-related effects. The three main model families that we study come from different geographical regions—the US, China, and Europe, respectively—which may result in differences in how they handle politically sensitive topics or interpret specific cultural references. Such a comparative approach aims to examine whether all models exhibit consistent patterns or if variations may emerge as a consequence of differences in technical capabilities, training data, and geopolitical development environment.

Each agent is initialized with varying degrees of behavioral context regarding the reference human user it simulates (e.g., political leaning, comment history, profile data) and prompted to reply to politically salient posts. We analyze LLMs' output along three axes: linguistic consistency, political leaning, and toxicity.

Our findings reveal clear patterns of a phenomenon that we refer to as *generative exaggeration*. Additional context improves linguistic coherence and ideological consistency with the reference users, but it simultaneously increases ideological extremism and verbal toxicity. Specifically, agents initialized with minimal context often fail to emulate real users faithfully. When provided with richer behavioral traces, the simulation improves in coherence but deteriorates in realism, producing stereotypical and exaggerated portrayals. This includes the overuse of partisan hashtags, emojis, and emotionally charged phrasing, as well as increased toxicity and partisan animosity. Importantly, this distortion is not symmetric: we find a consistent tendency to caricature right-leaning users more than left-leaning ones, though both are affected.

These results suggest that simulating political behavior with LLMs may not be a neutral process: LLM operators are tasked with solving the paradox that feeding more information and increasing computational performance may exaggerate certain traits and produce less realistic outputs, raising concerns about the use of LLMs in high-stakes contexts such as political communication and democratic deliberation.

This result adds to a growing literature using LLMs to simulate political personas [16–18], run polling experiments [19], and test moderation strategies in controlled settings [8, 20]. This body of work showed how models can exhibit emergent patterns over time, such as reinforcing bias or converging toward specific opinions [21–23], highlighting the importance of developing strategies to best align LLMs with users' behavior. However, the side effects of increasing simulation fidelity, such as the trade-off between staying close to a user's behavior and exaggerating some of its features, have not been studied in detail. While prior work has explored the biases of LLM-generated personas [24–27], we know little about how these models engage in political communication when simulating replies based on real user data.

Several studies have raised concerns about the tendency of LLMs to reflect or amplify societal biases [28, 29], including geopolitical bias [30, 31] and alignment with specific ideological narratives [32]. Other work has looked at the generation of hate speech in synthetic content [33], or at how model outputs can reinforce polarization in subtle ways [34, 35]. More generally, how LLMs shape digital conversations and structure discourse remains an open research question [36]. The present work shows how optimizing for fidelity along one user trait—political leaning—may spill over to unintended domains like verbal toxicity.

In this context, we make three main contributions. First, we provide empirical evidence on how LLMs simulate political users, showing systematic differences between the generated and original behavior. Second, we introduce the concept of generative exaggeration, a structural tendency of LLMs to amplify salient traits when simulating individuals. Third, we show that exaggeration affects political profiles asymmetrically. Our analysis suggests that these effects are not driven by prompt design but are a byproduct of model training and optimization. As a result, using LLMs as social agents—especially in sensitive contexts such as political communication—requires careful evaluation of the structural biases these systems may introduce.

Results and Discussion

To assess differences between human users and LLM, we analyze the behavior of LLM-based agents simulating politically active users on social media. The objective is to quantify the extent to which large language models replicate—and distort—human behavior in online political discourse.

To this end, we use a public dataset of over 21 million interactions on X (formerly Twitter) related to the 2024 U.S. presidential election [37]. This corpus enables the reconstruction of user profiles in a real-world context where political identity is salient and interactions are high-stakes.

We focus on tweet–reply interactions authored by 1,186 users with at least 50 prior tweets. This threshold ensures a sufficient behavioral signal to estimate political leaning and reliably capture individual linguistic style, while maintaining a large enough sample size for comparison. Each agent is thus prompted to respond to the same tweet as the original user, enabling direct tweet-level comparisons across multiple behavioral dimensions, including lexical diversity, ideological consistency, and toxicity of language.

To enable comparisons across the ideological spectrum, each user is assigned a political leaning score. We compute this score by applying a stance classifier to at least 50 tweets per user, labeling each message as pro-Democrat, pro-Republican, or neutral. Then, we assign a numerical value to each label (+1 for Republican, −1 for Democrat, 0 for Neutral) and average these values across the user's tweets. Based on the resulting score, users are categorized as Democrat, Neutral, or Republican. User- and LLM-generated responses are classified through the same procedure. Further methodological details are provided in the Methods section.

Agents are tested under two initialization conditions. In the Zero Shot setting, the model receives only the user's inferred political leaning. In the Few Shot setting, the model is provided with the user's nickname, bio, and a sample of their past tweets. The specific prompting strategies used in each case are described in the Methods section.

Our analysis is guided by four research questions, which we tackle in the following sections:

1. **Lexical Realism.** *Do LLM agents differ from humans—and from each other—in expressive style when simulating political users?*
2. **Ideological Consistency.** *How accurately do LLMs reproduce users' political leanings, and how does fidelity vary by prompt conditioning?*
3. **Toxicity Amplification.** *Do LLMs generate more toxic content than humans, and how does this depend on prompting strategy and political identity?*
4. **Generative Exaggeration.** *Do LLMs systematically amplify salient user traits, and under what conditions does this distortion emerge?*

Lexical Realism and Stylization

We start our analysis by assessing the ability of LLM-based agents to emulate human behavior at a linguistic level. First, we examine potential differences in length between tweets written by humans and by agents. This is an important consideration, as lexical measures are inherently sensitive to the length of the input text. Our results show that, although no model perfectly reproduces the distribution observed in human-authored tweets, the lengths of generated tweets generally fall within a plausible range, typically on the order of 10^2 characters, with the majority remaining under Twitter's 280-character constraint. However, we also identify a small fraction of anomalous tweets that exceed the maximum character length. These tweets are most often generated by the smaller models and are statistical outliers. Consequently, these instances were excluded from all subsequent lexical analyses. The tweet length distributions are displayed in full in Supplementary Figure S4, while a detailed breakdown of all removed tweets is reported in Supplementary Table S1.

We evaluate lexical diversity by computing the Type-Token ratio (TTR), which quantifies the proportion of unique words (types) to the total number of words (tokens) in the text. This measure allows for a straightforward interpretation: a higher TTR reflects a more varied vocabulary, while a lower TTR indicates frequent repetition of the same words. TTR-based measures are well-established in the literature as indicators of lexical variation and textual complexity [38–43]. In particular, we focus on the LogTTR, or Herdan's C [44–46], a variant of the standard TTR that is more robust to varying text lengths. The detailed formulation of this measure, as well as all text preprocessing steps performed, are reported in Methods.

In Fig. 1, we show how lexical diversity, measured using LogTTR, evolves as tweets are sequentially added to a growing corpus. The analysis is presented separately for each model and political leaning to assess any potential difference. By construction, all resulting curves are monotonically decreasing, since adding more text generally increases the number of unique words at a slower rate than the total number. Hence, in this analysis, we are interested in comparing how these curves decrease relative to one another across models and political leanings, as well as against human-generated tweets.

The panel in the top-right corner shows the LogTTR curves of human tweets alone, grouped by political leaning. Democrat and Republican users show similar levels of lexical diversity. In contrast, Neutral users tend to have higher diversity. This likely reflects the fact that non-partisan content covers a wider range of topics and vocabulary, whereas partisan tweets may often repeat similar terms, slogans, or rhetorical patterns.

Comparing human tweets with LLM-generated replies, we observe that all models tend to produce higher LogTTR scores than humans when considering small tweet samples, particularly for tweets inferred as Republican- or Democrat-leaning. However, this trend reverses when the number of examined tweets increases, as model-generated content begins to exhibit lower LogTTR values compared to human-authored tweets. This pattern indicates that although LLMs appear lexically rich over a small number of tweets, their vocabulary becomes increasingly repetitive as a larger corpus is taken into account.

The curves exhibit different end points due to the fact that a model may generate a tweet that reflects a political leaning different from that of its ground truth counterpart. In several cases, models show greater deviation from the ground truth when simulating Neutral users, though the extent varies across model families and prompting strategies. This political leaning category is underrepresented in the generated data due to the fact that most political shifts occur within it.

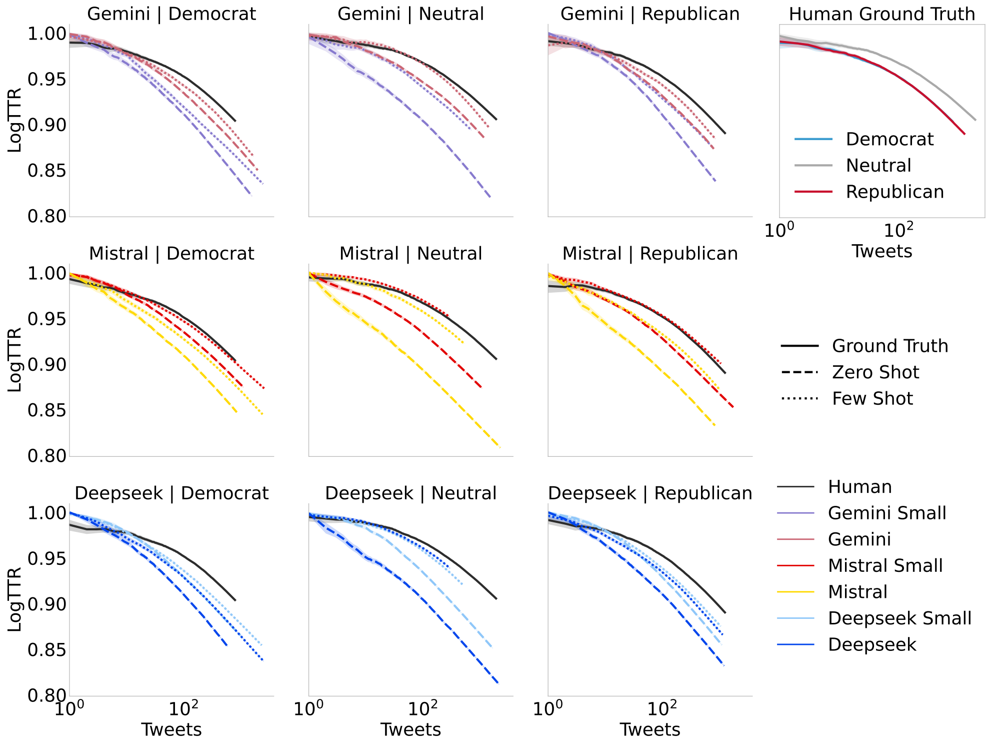


Fig. 1: LogTTR in the case of human and model-generated tweets, across different LLMs (rows) and political leanings (columns). Each panel shows how lexical diversity (LogTTR) evolves as more tweets are added (log-scaled). The curves represent the average over 100 simulations, each based on a different random ordering of tweets, to smooth variability and highlight general trends. Shaded areas indicate 95% confidence intervals, computed from 1000 bootstrap resamples. The Few Shot strategy appears to help models better approximate real human lexical behavior, bringing their output closer to human-authored content in terms of lexical diversity. In contrast, the top-right panel presents the LogTTR behavior of human-generated tweets exclusively, conditioned by political leaning. This analysis reveals that Democrat and Republican users exhibit comparable lexical diversity trends, while Neutral users demonstrate higher diversity.

Initialization strategy and model size both affect the lexical diversity manifested by agents. The Few Shot strategy, in particular, shifts the LogTTR curves of agent-generated content closer to those of real users. When comparing model sizes, the large variant of Gemini better approximates human lexical diversity than its smaller counterpart. In contrast, for the Mistral and DeepSeek families, the smaller models produce outputs more consistent with human behavior. Among all configurations, Mistral Small with Few Shot initialization produces the LogTTR curves that most closely match those of human users across all political leanings.

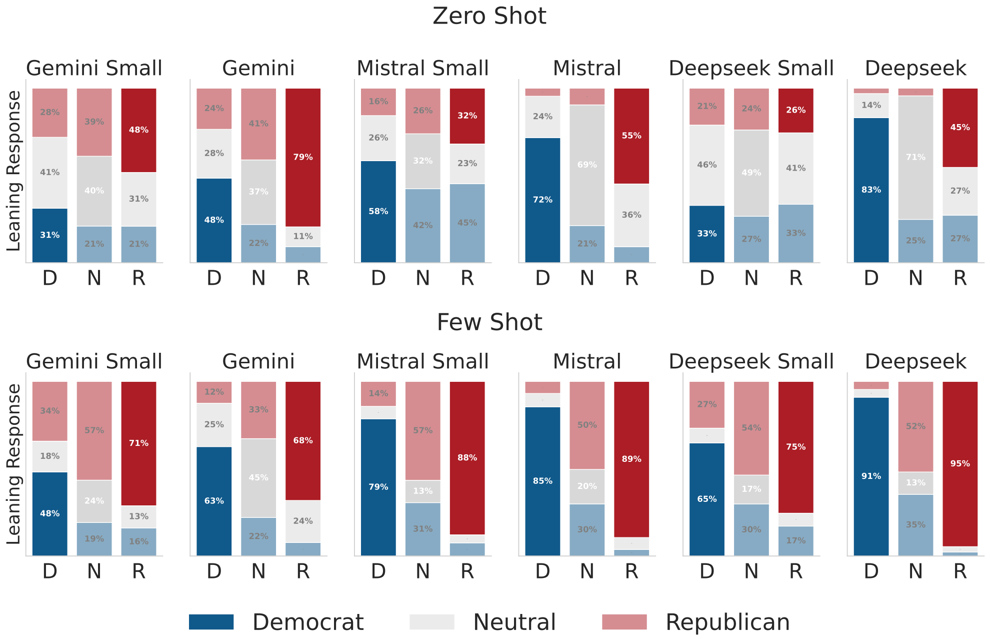


Fig. 2: For each user simulated via an LLM-based agent, we evaluate the political leaning of the generated tweets (y-axis) with respect to the user's original leaning class (x-axis), and report the percentage of tweets that are consistent with the initialized political leaning. When models are initialized in a Zero Shot manner—using only the user's political leaning—they tend to produce politically neutral outputs, lacking fidelity to the intended ideological stance. However, when Few Shot prompting is used to provide contextual examples, models more accurately embody the target political identity, generating comments that better reflect the intended leaning. This increased alignment comes at the cost of decreased neutrality for users labeled as politically neutral.

Repeating the analysis using the standard TTR metric yields results that are consistent with those obtained from LogTTR (see Supplementary Figure S4). A summary of the corresponding quantitative analysis of the texts is provided in Supplementary Table S2.

These findings suggest that, while current LLMs can reproduce surface-level lexical variation, their outputs are still influenced by internal priors that diverge from those observed in human behavior. This indicates inherent constraints in the capacity of LLMs to faithfully simulate real users, even under optimized initialization strategies.

Ideological Bias and Consistency

In this section, we assess whether agent responses align ideologically with the users they are meant to simulate. Following previous work [47, 48], we analyze the inferred political leaning of the generated replies. Agents are first grouped according to the political leaning of their target user—Democrat, Neutral, or Republican. We then apply an automated classifier [49] to label each generated reply and compute the conditional probability of producing a reply with a given leaning, given the emulated user's profile. Note that in the Zero Shot condition, agents are explicitly given the user's political leaning, whereas in the Few Shot condition, this information is not directly provided unless included in the user's tweets or bio.

Figure 2 shows the distribution of political leanings in tweets generated by agents, grouped by the political leaning of the corresponding users. We observe a clear difference across prompting strategies. In the Zero Shot setting, where agents are only prompted with the user's political leaning, no clear pattern emerges. Most agents tend to produce replies across the whole political spectrum, regardless of the political leaning they were provided at initialization. In contrast, Few Shot prompting, which is based on a user's tweet history, results in responses more consistent with the user's original stance, but at the cost of decreased neutrality. Notably, users originally labeled as Neutral are more often emulated as more politically aligned than they actually are, suggesting a context-driven distortion in how LLMs capture political identity.

For instance, when prompted to simulate a Republican-leaning user, DeepSeek generates ideologically aligned responses in approximately 45% of cases in the Zero Shot condition. However, this proportion rises to 95% in the Few Shot setting, indicating an improved consistency with the reference ideology. At the same time, when focusing on users labeled as Neutral, we observe that the proportion of generated tweets drops from 71% in the Zero Shot setting to 13% in the Few Shot setting. This shift highlights a tendency of LLMs to drift toward polarized outputs even when exposed to minimal ideological cues. An exception is the larger Gemini model, which exhibits a high proportion of partisan comments even in the Zero Shot setting, as well as a sizable presence of neutral comments in the Few Shot setting.

Figure 3 illustrates the models' 'ideological consistency', a metric we introduce to quantify the coherence between an individual's expressed opinions (through comments) and their known ideological leaning. Specifically, it measures how closely the political leaning of generated replies aligns with the user's political leaning. To compute this, we first measure, separately for humans and agents, how closely each message aligns with the user's estimated political identity. We then average this alignment across all messages and users to obtain the 'ideological consistency loss' $\mathcal{L}$. Finally, we define $\mathcal{C}^A = 1 - \mathcal{L}^A$ as the ideological consistency of agents and $\mathcal{C}^H = 1 - \mathcal{L}^H$ of humans. A formal definition of these metrics is provided in Methods.

The values of ideological consistency range from 0 to +1. A higher value indicates that the comments are more likely to reflect the true political leaning of the user or agent, demonstrating strong alignment between expressed opinions and ideological stance. Conversely, a value closer to zero suggests little or no alignment.

In the Zero-Shot setting, the ideological consistency of most agents simulating Republican users is lower than that observed in humans, suggesting that their responses are less ideologically aligned compared to their human counterparts. The larger variants of Gemini and Mistral are exceptions to this tendency. For Democrat and Neutral users, instead, we observe a larger variation across models. Notably, agents

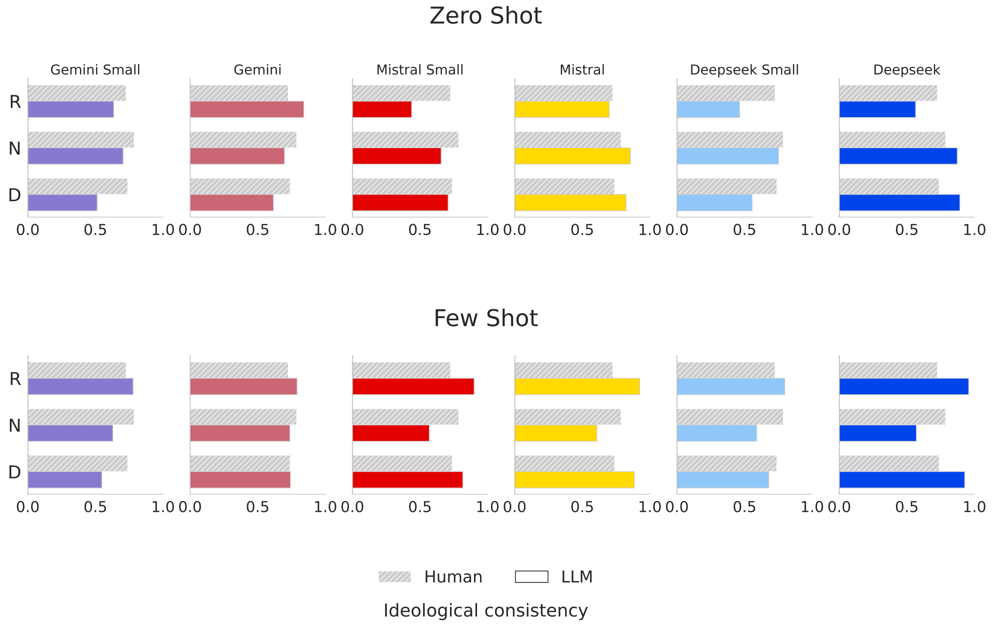


Fig. 3: Ideological consistency for each model for the two initialization strategies (Zero Shot or Few Shot). The metric ranges from 0 to +1 and measures the coherence between the agent's initial political leaning and that of the generated comments. See Eq. (8) in Methods for its formal definition. A value close to 1 indicates a caricatured ideological portrayal by LLMs and that all responses contain political content expressing the subject's political leaning. Conversely, a value close to 0 results from agents employing a less politically overt tone than their human counterparts. Overall, agents generate outputs that more accurately reflect the user's stance in the Few Shot setting, though there are exceptions, such as DeepSeek or Mistral.

modeling Neutral users are generally characterized by an ideological consistency similar to that observed in humans. Mistral stands out as the model that most faithfully replicates the political stance of users across the spectrum.

When evaluating agents initialized with the Few Shot procedure, we find notable differences from the Zero Shot setting across all models, except for the two Gemini variants. Overall, adding a user's comment history to the prompt tends to increase the ideological consistency of agent responses with the original users, as shown in Fig. 2. However, this increase in alignment is accompanied by other changes that may reduce the behavioral consistency of the outputs. The generally high ideological consistency for both Democrat- and Republican-leaning users suggests that agents produce tweets that align more closely with the model's own political leaning than the original users do. In this way, the models generate a caricatured representation of the user's political leaning within the debate. This is particularly evident for the larger Mistral and DeepSeek models, both displaying a substantially different behavior when

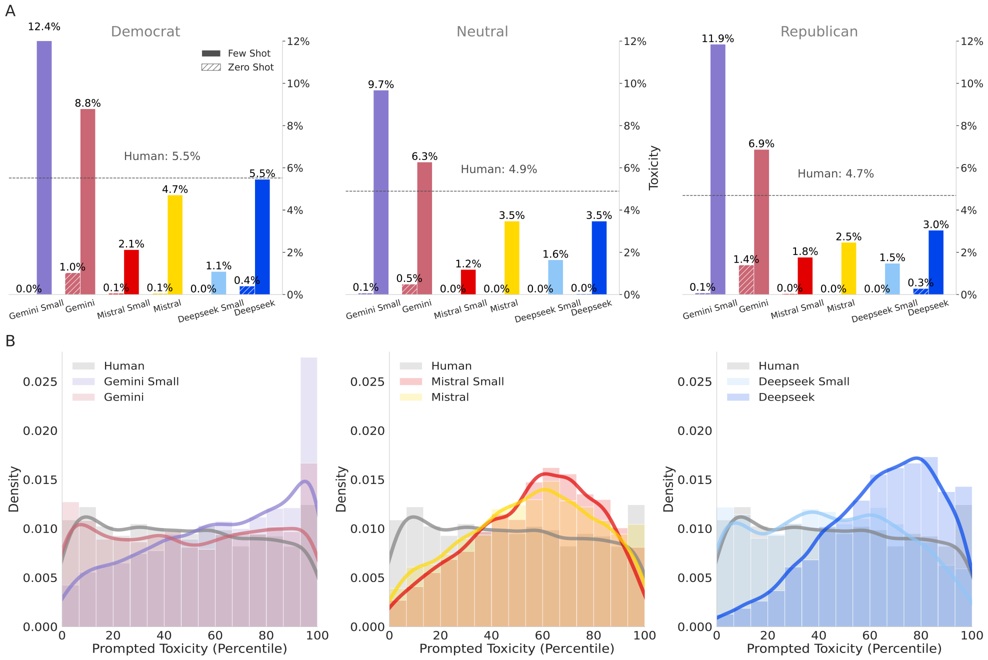


Fig. 4: (A) Proportion of responses with toxicity score > 0.6, grouped by LLM, initialization method (Zero Shot and Few Shot), and political leaning. Providing contextual comments results in higher toxic output, especially with models like Gemma and Gemini, both developed by Google, which exceed the human baseline average of 5%. (B) Distribution of toxicity percentiles for synthetic tweets generated by different models. Each tweet's toxicity percentile is calculated relative to the toxicity levels of the 30 comments in the prompt, indicating where the tweet's toxicity ranks within that reference set. While Mistral-based models tend to produce a roughly Gaussian distribution, Google's models exhibit a relatively uniform shape with a sharp peak at the 100th percentile, suggesting a tendency to exceed the toxicity level present in the prompt.

compared to the Zero Shot case. In contrast, the larger Gemini model produces the agents that best reproduce the original users' political output in this setting.

Toxicity

To quantify the prevalence of toxic or harmful content in generated tweets, we annotate each reply using Perspective API [50], a widely adopted classifier for detecting toxic language in online content [51]. The API defines toxicity as "a rude, disrespectful, or unreasonable comment likely to make someone leave a discussion", and assigns a toxicity score ranging from 0 (non-toxic) to 1 (highly toxic). Specifically, we compute the proportion of tweets exceeding a toxicity score of 0.6, a threshold commonly adopted in prior work on online toxicity to label a message as 'toxic' [51].

Leveraging these annotations, we assess both the prevalence and severity of toxic language in human replies and in their LLM-generated counterparts, evaluating the degree to which such behavioral traits are faithfully preserved or distorted in the simulation. We apply this metric across all agent replies, comparing model outputs to human baselines. Additionally, we group the results by the political leaning class used to prompt the agents, allowing us to examine whether toxicity levels vary systematically across the political spectrum.

As shown in Fig. 4A, the prompting strategy (Zero or Few Shot) appears to play a crucial role in the toxicity levels exhibited by the agents. When agents are initialized using the Zero Shot procedure, they produce a small fraction of toxic replies across all LLMs under evaluation. Users, in turn, produce approximately 5% toxic comments. When agents are provided with a set of user comments, in the Few Shot setting, toxicity does emerge. This is particularly evident for Gemini models, which, in this setting, frequently generate toxic content, even exceeding human reference levels.

Toxicity can appear in LLM agents when it exists in the tweets used as prompts, even if it is not the typical behavior of the models. Models generally respect safety constraints in the Zero Shot setting, but Few Shot prompting—which adds more user context—can weaken these controls and allow harmful content. This trade-off poses a challenge for using LLMs in social simulations: improving behavioral accuracy may reduce safety measures.

Building on these findings, to assess whether LLM agents reflect or amplify the toxicity levels present in their prompting context, we compare the toxicity of generated responses with the percentile distribution of toxicity in the user's historical tweets (i.e., the Few Shot prompt strategy). This approach allows us to evaluate whether agents reproduce the observed variability in human behavior or exhibit deviations. Deviations from the user's baselines, such as consistent overproduction of toxic content relative to the provided tweets, would signal the emergence of systematic exaggeration.

Formally, consider a tweet with toxicity score t and a set of M reference tweets with toxicity scores $\{r_j\}_{j=1}^{M}$. The percentile rank $P(t)$ is defined as the proportion of reference tweets with toxicity less than or equal to t:

$$P(t) = \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}\big(r_j \leq t\big) \tag{1}$$

where $\mathbf{1}(\cdot)$ denotes the indicator function. In our specific setup, we set $M = 30$, which corresponds to the number of tweets provided to each agent for the Few Shot initialization.

As shown in Fig. 4B, nearly all models tend to *overshoot* the reference toxicity range of the users they simulate: the distribution of toxicity scores in the generated replies has a center of mass above the 50th percentile of the prompt distribution in most cases. This effect is particularly pronounced in the larger model variants, which often generate outputs more toxic than the tweet samples they were initialized with. For instance, larger versions of Mistral and DeepSeek produce left-skewed distributions, indicating a consistent upward shift in toxicity. Gemini, in contrast, displays a sharply peaked distribution—especially in its smaller variant—suggesting that the generated replies cluster around the most toxic subset of the prompt tweets. This pattern reveals a generative bias toward amplifying the more extreme or emotionally charged elements of the input, rather than sampling proportionally from the full distribution of prior user behavior.

The Style of Generative Exaggeration

Our findings suggest that LLMs do not faithfully reproduce the users they are meant to simulate, across several key behavioral attributes such as ideology and toxic language. Instead, they tend to amplify specific prominent traits, especially in the Few Shot setting, resulting in a systematic distortion of user behavior. This reflects a form of *generative exaggeration*, in which the agent captures superficial markers of identity while misrepresenting the underlying behavioral profile and creating a caricatured portrayal. This behavior is not incidental. It reflects a tendency of models to prioritize linguistic salience and consistency over contextual depth. When models are conditioned on partisan cues, they do not infer ideology—they generate responses based on prominent patterns in the input data. Hence, fidelity collapses into caricature.

We now aim to better characterize this phenomenon, shifting our attention toward the use of emojis and hashtags in the agent-generated replies. This is particularly relevant in the context of political discourse, as these elements are typically employed as markers of ideological partisanship, especially in a polarized setting.

Further characterizing the results presented in Fig. 3, human-authored tweets contain emojis and hashtags relatively rarely—87.47% and 96.29% of tweets, respectively, omit them. In contrast, agents include these elements much more frequently, especially in the Few Shot setting. For instance, Mistral-generated replies contain emojis and hashtags in over 49% of cases. See Supplementary Table S2 for a full breakdown.

To evaluate the extent of this amplification, Fig. 5 shows the proportion of human- and agent-generated tweets containing each emoji or hashtag present in both sets. In the figure, a ratio of 1 indicates an equal number of occurrences in both sets, while a ratio greater (smaller) than 1 is a marker of LLM overrepresentation (underrepresentation). We observe that overall, emoji use is markedly exaggerated, especially for those linked to stereotypical political identities. For example, the rainbow emoji—typically associated with progressive viewpoints—appears nearly 20 times more often in DeepSeek-generated tweets than in authentic human replies. A comparable inflation is observed in hashtag usage. While some ideologically charged hashtags (e.g., *#kamalaistheproblem* or *#oldmantrump*) are slightly underrepresented, others are substantially amplified. For instance, hashtags like *#kamalaharris2024*, *#MAGA*, and *#MAGA2024* appear up to 10-15 times more frequently in LLM outputs than in real user content. Although Fig. 5 focuses on DeepSeek in the Few Shot setting, comparable behaviors are found across all evaluated models (see Supplementary Figures S5 and S6 for hashtags and emojis, respectively).

The over-production we observe is systematic, not random. During training, LLMs minimise next-token error by giving extra weight to high-salience ideological tokens—slogans, labels, polarising phrases—because those tokens strongly predict surrounding text. The generated output therefore looks ideologically coherent, yet it collapses rich political positions into a few over-used symbols. If such models take over content moderation or deliberation, this skew becomes part of the decision pipeline: instead

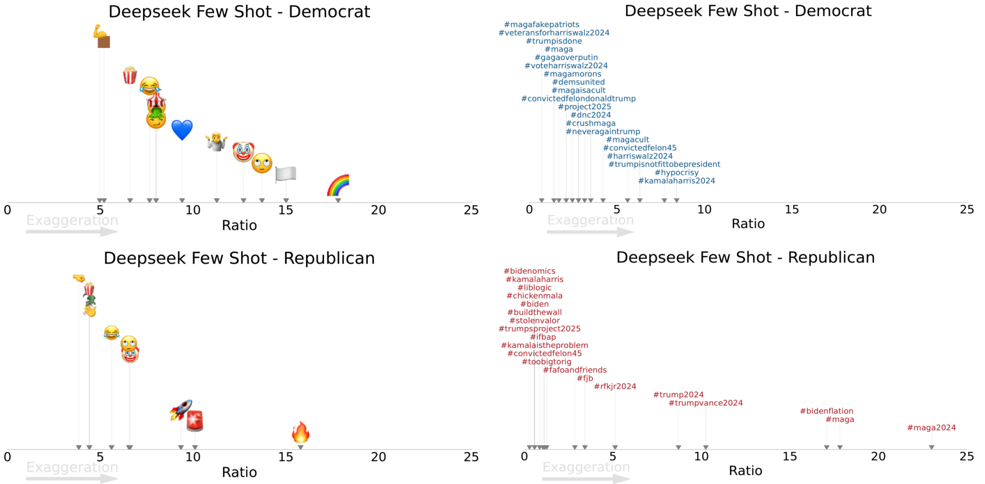


Fig. 5: We analyze the use of emojis and hashtags in tweets classified as Democratic or Republican, comparing human and LLM-generated content. The arrow labeled "Exaggeration" starts on the x-axis at a ratio of 1 and marks the beginning of LLM overrepresentation. In Panel (A), each point represents an emoji used either by humans or by models. The x-axis shows the ratio between the relative frequencies—defined as the fraction of tweets containing the emoji—of LLM-generated versus human-generated content, computed separately for Democratic and Republican tweets. This visualization highlights emojis that are disproportionately used by the model compared to humans. Models tend to overuse expressive and stereotypical emojis, such as the clown face, laughing face, and confused face. Notably, the rainbow (associated with Democratic discourse) is more frequent in model-generated Democratic tweets, while the fire or the rocket appear more in Republican ones. Panel (B) replicates the same analysis for hashtags: the x-axis again represents the ratio between model and human relative frequencies. The results reveal a consistent overuse of politically charged hashtags by the model across both political alignments, with a slight overrepresentation in Republican tweets.

of merely routing discourse, the agents reshape it by amplifying loud cues and muting low-frequency nuance.

Conclusions

This study investigated how large language models behave when tasked with simulating users engaged in online political discourse. We constructed over 1,000 synthetic agents modeled on real users involved in the debate on X surrounding the 2024 U.S. presidential election, prompting them to reply to politically salient content under controlled experimental conditions. Agents were initialized either with a Zero Shot procedure, receiving only the user's estimated political leaning, or a Few Shot procedure, where they are provided with the user's bio and message history.

Across four key dimensions—lexical diversity, ideological consistency, toxicity, and stylistic exaggeration—our results show a systematic pattern: LLMs reshape user behavior in ways that reflect structural biases, rather than simply mirroring it. In the Zero Shot setting, responses are distributed uniformly across the political spectrum, regardless of the prompted ideological leaning. This suggests that indicating a political leaning alone is insufficient to steer the model's ideological output. When behavioral priors are introduced, agents become more ideologically consistent but also more extreme. This behavior is denoted by intensified partisan tones, amplified markers of political identity, such as emojis and hashtags, and, in several cases, increased toxicity.

We refer to this observed tendency as *generative exaggeration*: a structural distortion wherein agents only capture and amplify salient features of the users they are meant to simulate, especially ideological cues. This distortion is neither neutral nor uniform across the political spectrum. Crucially, this does not reflect a failure of LLMs to create ideologically consistent personas. Rather, it entails the generation of agents that are only capable of displaying stereotypical or caricatured portrayals of their human counterparts, generating unprompted but systematic behaviors like heightened toxicity.

These findings carry direct implications for the deployment of LLMs as social agents. Whether in moderation pipelines, deliberative systems, or synthetic media generation, these models risk introducing systematic biases, potentially reinforcing polarization and presenting ideological caricatures as ordinary behavior. The observed relationship between ideology and toxicity suggests that alignment protocols built around safety may inadvertently encode political valence.

Limitations of our analysis should also be acknowledged. Our study is limited to the U.S. political discourse and single-turn interactions. Future work should explore multi-turn dialogue, cross-linguistic and cultural generalizability, and longitudinal effects in live environments. Moreover, standard classifiers used for toxicity and ideological labeling may themselves carry biases. We should also note that our analysis focused solely on political debate, which may have influenced the ideological outputs of the models. Including users' general tweet histories might have led to different results. However, given our aim of simulating user behavior specifically within political discourse, as well as the current limitations in retrieving complete longitudinal user activity, we believe that focusing on politically relevant content is a well-justified approach.

Ultimately, the emerging use of LLMs as proxies for human behavior requires a new methodological posture. Rather than asking whether models replicate surface traits, we must interrogate how they misrepresent structure. Generative exaggeration, as we show, is a byproduct of systems optimized for salience over subtlety. Any serious attempt to deploy LLMs in socially meaningful contexts must begin by accounting for this epistemic drift.

Methods

Agent Initialization

Modeling human behavior through LLM agents usually involves conditioning the model on specific user attributes to guide its responses, such as demographics or political leaning [17, 23, 52, 53]. By embedding such information in the prompt, the LLM is instructed to produce outputs that align with the provided persona.

In this work, we generate synthetic agents using six language models drawn from three model families: Gemini, Mistral, and DeepSeek. For each family, we include both a smaller and a larger variant: Gemini 2.0 Flash and Gemma 3 (4.3B) for Gemini, Mistral (7B) and Mistral (123B) for Mistral, and DeepSeek V2 (7B) and DeepSeek V3 (671B) for DeepSeek.

Our experiments focus on dyadic exchanges, prompting LLMs to reply to real tweets and comparing their responses to human replies. To this end, we initialize each LLM agent using two distinct approaches:

- **Zero Shot:** The model is provided with the estimated political leaning score of the user of the to-be-modeled users.
- **Few Shot:** The model is provided with 30[1] tweets written by the user it is being modeled after, as well as the user's bio.

The second approach, which incorporates users' past comments as input, enables us to measure the extent to which political leaning is inferred directly from the messages without being explicitly provided, and consequently to evaluate how this inference differs from the model's Zero Shot understanding of that leaning. The exact prompts used for the Zero Shot and Few Shot initializations are reported in Fig. 6 and Fig. 7, respectively.

[1] this threshold was chosen to make sure that the prompt was not too long nor truncated.

Zero Shot Prompt

```
### Your Profile:
You are a Twitter user with a {leaning} political leaning:
- Values around -1 indicate left-leaning views
   (e.g., Kamala Harris or Joe Biden support).
- Values near 0 indicate a neutral or independent stance.
- Values close to 1 indicate right-leaning views
   (e.g., Trump support).

### Conversation Context:
Here is the conversation thread so far:
{thread}

### Task:
Write a tweet to reply according to your profile
(without including any context or explanations),
with not more than 100 characters.
```

Fig. 6: Zero Shot Prompt: Simulates a user based solely on political leaning. No specific identity or style data is provided.

Few Shot Prompt

```
### Your Data:
- **Your Usernames:** {usernames}
    (These are the usernames by which you are known on Twitter.
     Use them as your identity when responding.)
- **Your Bios:** {biographies}
    (These reflect how you present yourself, your ideology,
     and your values. Refer to these to guide your tone,
     priorities, and responses.)
- **Your Tweets:** {tweets}
    (These demonstrate your personality, tone,
     topics of interest, and typical style of communication.)

### Conversation Context:
Here is the conversation thread so far:
{thread}

### Task:
Write a tweet to reply according to your profile
(without including any context or explanations),
with not more than 100 characters.
```

Fig. 7: Few Shot Prompt: Simulates a user using rich user-specific data, including usernames, bios, and prior tweets.

Political Leaning Estimation

To estimate the political leaning of a user, we consider at least 50 comments previously posted by the user. Hence, we apply the method described in [49], which classifies the stance of a message as supportive of the Democratic Party, the Republican Party, or as neutral. Accordingly, we assign each message a numerical score: +1 if pro-Republican, −1 if pro-Democratic, and 0 if neutral. A message is considered neutral if it does not contain phrases explicitly expressing support for either Trump and the Republican party, or Biden, Harris, and the Democratic party. Finally, for each user i, we compute their political leaning score L_i as the arithmetic mean of the individual message scores:

$$L_i = \frac{1}{M_i} \sum_{j=1}^{M_i} s_{ij} \tag{2}$$

where $s_{ij} \in \{-1, 0, +1\}$ is the score assigned to the j-th message of user i, and $M_i = 50$ is the total number of evaluated messages. The resulting L_i provides a continuous measure of political leaning, ranging from -1 (strongly Democratic-leaning) to $+1$ (strongly Republican-leaning), with values around 0 indicating ideological neutrality.

To evaluate how ideologically faithful each agent's replies are to its human counterparts, we define a metric called ideological consistency loss. This metric captures the degree to which a given response aligns with the overall ideological leaning of the user, as inferred from their past comments (L_i). For each user i and reply k, we define it as

$$\ell(C_i, s_{ik}) = \frac{|C_i - s_{ik}|}{2} \tag{3}$$

where C_i is the binned leaning L_i of user i:

$$C_i = \begin{cases} -1 & \text{if } L_i < -0.25 \quad \text{(Left-leaning)} \\ 0 & \text{if } -0.25 \leq L_i \leq 0.25 \quad \text{(Moderate/Neutral)} \\ +1 & \text{if } L_i > 0.25 \quad \text{(Right-leaning)}\,. \end{cases} \tag{4}$$

The denominator of Eq. (3) is used to normalize the squared ideological distance between comments and users in the range $[0, 1]$, since the maximum possible distance between ideological scores in the interval $[-1, 1]$ is equal to 2. This normalization facilitates the comparison and visualization of differences across ideological classes.

To quantify ideological consistency, we separately compute Eq. (3) for all human-generated replies s_{ik}^H authored by user i (where the superscript H denotes human), and for all synthetic replies s_{ik}^A produced by the LLM agent simulating user i (with A denoting agent).

Each synthetic reply is generated using the same tweet prompt as the corresponding human reply, ensuring a controlled comparison. As detailed previously, agents are initialized either in a Zero Shot or Few Shot setting, depending on whether they are provided only with metadata (e.g., political leaning) or with additional contextual information such as tweet history. This procedure allows us to isolate the effect of prompting on ideological reproduction.

By computing ideological consistency loss separately for humans and agents, we obtain a user-level measure of how accurately each agent preserves the political leaning of the person it is designed to emulate. This metric enables us to evaluate both average performance across the population and model-specific deviations in consistency.

Next, for each $C \in \{-1, 0, +1\}$, we average the ideological consistency loss over all replies $R = \sum_{i=1}^{N_C} R_i$ left by the N_C users with $C_i = C$, and their corresponding agents:

$$\mathcal{L}_C^H = \frac{1}{R} \sum_{i=1}^{N_C} \sum_{k=1}^{R_i} \ell(C, s_{ik}^H) \tag{5}$$

$$\mathcal{L}_C^A = \frac{1}{R} \sum_{i=1}^{N_C} \sum_{k=1}^{R_i} \ell(C, s_{ik}^A) \tag{6}$$

Therefore, $\mathcal{L}^H$ captures the average ideological consistency loss of all humans, while $\mathcal{L}^A$ measures it for all agents. We note that, since each human has a corresponding LLM agent, and for every human reply there is an associated LLM-generated reply, the number of replies R and the number of users N_C are identical for both $\mathcal{L}^H$ and $\mathcal{L}^A$. We apply this procedure separately for each LLM, calculating the consistency between the comments made by a human or an agent and their corresponding leaning

$$\mathcal{C}_C^H = 1 - \mathcal{L}_C^H \tag{7}$$

$$\mathcal{C}_C^A = 1 - \mathcal{L}_C^A \tag{8}$$

Lexical Diversity Analysis

In this work, we assess lexical diversity using two metrics: the standard Type-Token ratio (TTR) and a variant known as LogTTR.

TTR-based metrics inherently rely on the Bag-of-Words assumption, treating a document as an unordered collection of words. This abstraction enables a focus on vocabulary richness and frequency distribution while disregarding syntactic and sequential information, which is less relevant for this type of lexical analysis. Unlike stance detection, the goal of this lexical approach is not to infer the user's ideological stance or reaction to the tweet being replied to. Consequently, it is not well suited for capturing semantic distinctions between tweets with similar vocabulary. For example, *"I'll vote for X"* and *"I'll not vote for X"* would be considered highly similar in lexical terms, despite expressing opposing meanings. Because TTR-based metrics rely on the accurate identification of word tokens, effective tokenization is a crucial preprocessing step. For this purpose, we adopt the SpaCy tokenizer [54] with the `en_core_web_sm` language model, which provides robust, linguistically informed tokenization.

In our preprocessing pipeline, we remove stop words, punctuation, hashtags, URLs, and email addresses to retain only semantically meaningful content. Additionally, named entities composed of multiple words are treated as single tokens—for example, the sentence *"New York is a #busy city!"* is tokenized as `["New York", "is", "a", "city"]` rather than `["New", "York", "is", "a", "city"]`. Furthermore, different denominations or lexical variants of the same underlying entity (e.g., *"U.S."* vs. *"United States"*) are treated as distinct tokens to more accurately reflect differences in how entities are expressed and to quantify lexical diversity better.

After all text preprocessing steps, for a given text document d_i with τ_i unique word types and ω_i total word tokens, we compute the LogTTR as:

$$\mathrm{LogTTR}(d_i) = \frac{\log(\tau_i + \alpha)}{\log(\omega_i + \alpha)}, \tag{9}$$

In this formulation, we introduce a smoothing term $\alpha = 1$ in both the numerator and denominator to ensure that the metric remains well-defined even for extremely short texts, such as tweets. In particular, it allows us to compute LogTTR for minimal-length documents, including cases where $\tau_i = \omega_i = 1$, thereby ensuring robustness and comparability across all tweets in our corpora.

Data availability

The X/Twitter data used for all our experiments has been made publicly available on a GitHub repository (https://github.com/sinking8/x-24-us-election) by the dataset authors [37]. Specifically, we use the version of the dataset corresponding to the latest repository commit as of January 15, 2025.

Author contribution

W.Q. supervised the research. J.N., M.E.P., and E.L. implemented the analysis pipeline and conducted the experiments. J.N., M.E.P., E.L., M.S., and M.C. designed the methodological framework and performed data interpretation. All authors contributed to writing and revising the manuscript.

Acknowledgements

We thank Geronimo Stilton and the Hypnotoad for inspiring the data analysis and the interpretation of the results. This work was supported by the SERICS project (PE00000014) under the NRRP MUR program funded by the European Union – NextGenerationEU; the DECODE project funded by the Presidency of the Council of Ministers (Italy); and the MUSMA project (PRIN 2022, CUP G53D23002930006) funded by the Italian Ministry of University and Research (MUR) under the EU NextGenerationEU initiative, Mission 4, Component 2, Investment 1.1.

References

[1] Debenedetti, E., Zhang, J., Balunovic, M., Beurer-Kellner, L., Fischer, M., Tramèr, F.: Agentdojo: A dynamic environment to evaluate prompt injection

attacks and defenses for llm agents. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024)

[2] Boskabadi, M.R., Cao, Y., Khadem, B., Clements, W., Gerek, Z.N., Reuthe, E., Sivaram, A., Savoie, C.J., Mansouri, S.S.: Industrial agentic ai and generative modeling in complex systems. Current Opinion in Chemical Engineering **48**, 101150 (2025)

[3] Shen, Y., Song, K., Tan, X., Li, D., Lu, W., Zhuang, Y.: Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. Advances in Neural Information Processing Systems **36**, 38154–38180 (2023)

[4] Goodell, A.J., Chu, S.N., Rouholiman, D., Chu, L.F.: Large language model agents can use tools to perform clinical calculations. npj Digital Medicine **8**(1), 163 (2025)

[5] M. Bran, A., Cox, S., Schilter, O., Baldassari, C., White, A.D., Schwaller, P.: Augmenting large language models with chemistry tools. Nature Machine Intelligence **6**(5), 525–535 (2024)

[6] Kim, S., Yu, Y., Seo, H.: Artificial intelligence orchestration for text-based ultrasonic simulation via self-review by multi-large language model agents. Scientific Reports **15**(1), 12474 (2025)

[7] Coppolillo, E., Cinus, F., Minici, M., Bonchi, F., Manco, G.: Engagement-driven content generation with large language models. arXiv preprint arXiv:2411.13187 (2024)

[8] Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual Acm Symposium on User Interface Software and Technology, pp. 1–22 (2023)

[9] Epstein, J.M., Axtell, R.: Growing Artificial Societies: Social Science from the Bottom Up. Brookings Institution Press, ??? (1996)

[10] Conte, R., Edmonds, B., Moss, S., Sawyer, R.K.: Sociology and social theory in agent based social simulation: A symposium. Computational & Mathematical Organization Theory **7**(3), 183–205 (2001)

[11] Møller, A.G., Romero, D.M., Jurgens, D., Aiello, L.M.: The impact of generative ai on social media: An experimental study. arXiv preprint arXiv:2506.14295 (2025)

[12] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proceedings of the national academy of sciences **118**(9), 2023301118 (2021)

[13] Di Martino, E., Galeazzi, A., Starnini, M., Quattrociocchi, W., Cinelli, M.: Characterizing the fragmentation of the social media ecosystem. arXiv preprint arXiv:2411.16826 (2024)

[14] Del Vicario, M., Bessi, A., Zollo, F., Petroni, F., Scala, A., Caldarelli, G., Stanley, H.E., Quattrociocchi, W.: The spreading of misinformation online. Proceedings of the national academy of Sciences **113**(3), 554–559 (2016)

[15] Donkers, T., Ziegler, J.: Understanding online polarization through human-agent interaction in a synthetic llm-based social network. In: Proceedings of the International AAAI Conference on Web and Social Media, vol. 19, pp. 457–478 (2025)

[16] Koley, G.: Salm: A multi-agent framework for language model-driven social network simulation. arXiv preprint arXiv:2505.09081 (2025)

[17] Rossetti, G., Stella, M., Cazabet, R., Abramski, K., Cau, E., Citraro, S., Failla, A., Improta, R., Morini, V., Pansanella, V.: Y social: an llm-powered social media digital twin. arXiv preprint arXiv:2408.00818 (2024)

[18] Cau, E., Failla, A., Rossetti, G.: Bots of a feather: Mixing biases in llms' opinion dynamics. In: International Conference on Complex Networks and Their Applications, pp. 166–176 (2024). Springer

[19] Holtdirk, T., Assenmacher, D., Bleier, A., Wagner, C.: Fine-tuning large language models to simulate german voting behaviour. Technical report, Center for Open Science (2024)

[20] Park, J.S., Popowski, L., Cai, C., Morris, M.R., Liang, P., Bernstein, M.S.: Social simulacra: Creating populated prototypes for social computing systems. In: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, pp. 1–18 (2022)

[21] Ahnert, G., Wurth, E., Strohmaier, M., Mata, J.: Simulating persuasive dialogues on meat reduction with generative agents. arXiv preprint arXiv:2504.04872 (2025)

[22] Coppolillo, E., Manco, G., Aiello, L.M.: Unmasking conversational bias in ai multiagent systems. arXiv preprint arXiv:2501.14844 (2025)

[23] Taubenfeld, A., Dover, Y., Reichart, R., Goldstein, A.: Systematic biases in llm simulations of debates. arXiv preprint arXiv:2402.04049 (2024)

[24] Cheng, M., Piccardi, T., Yang, D.: Compost: Characterizing and evaluating caricature in llm simulations. arXiv preprint arXiv:2310.11501 (2023)

[25] Liu, A., Diab, M., Fried, D.: Evaluating large language model biases in

persona-steered generation. In: Findings of the Association for Computational Linguistics ACL 2024, pp. 9832–9850. Association for Computational Linguistics, ??? (2024). https://doi.org/10.18653/v1/2024.findings-acl.586 . http://dx.doi.org/10.18653/v1/2024.findings-acl.586

[26] Li, A., Chen, H., Namkoong, H., Peng, T.: LLM generated persona is a promise with a catch (2025) 2503.16527 [cs.CL]

[27] Alipour, S., Sen, I., Samory, M., Mitra, T.: Robustness and confounders in the demographic alignment of llms with human perceptions of offensiveness. arXiv preprint arXiv:2411.08977 (2024)

[28] Hu, T., Kyrychenko, Y., Rathje, S., Collier, N., Linden, S., Roozenbeek, J.: Generative language models exhibit social identity biases. Nature Computational Science **5**(1), 65–75 (2025)

[29] Loru, E., Nudo, J., Di Marco, N., Cinelli, M., Quattrociocchi, W.: Decoding ai judgment: How llms assess news credibility and bias. arXiv preprint arXiv:2502.04426 (2025)

[30] Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., Johary, I., Mara, A.-C., Romero, R., Lijffijt, J., et al.: Large language models reflect the ideology of their creators. arXiv preprint arXiv:2410.18417 (2024)

[31] Noels, S., Bied, G., Buyl, M., Rogiers, A., Fettach, Y., Lijffijt, J., De Bie, T.: What large language models do not talk about: An empirical study of moderation and censorship practices. arXiv preprint arXiv:2504.03803 (2025)

[32] Chen, K., He, Z., Yan, J., Shi, T., Lerman, K.: How susceptible are large language models to ideological manipulation? arXiv preprint arXiv:2402.11725 (2024)

[33] Civelli, S., Bernardelle, P., Demartini, G.: The impact of persona-based political perspectives on hateful content detection. In: Companion Proceedings of the ACM on Web Conference 2025, pp. 1963–1968 (2025)

[34] Piao, J., Lu, Z., Gao, C., Xu, F., Santos, F.P., Li, Y., Evans, J.: Emergence of human-like polarization among large language model agents. arXiv preprint arXiv:2501.05171 (2025)

[35] Sharma, N., Liao, Q.V., Xiao, Z.: Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In: Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems, pp. 1–17 (2024)

[36] Lazovich, T.: Filter bubbles and affective polarization in user-personalized large language model outputs. In: Proceedings On, pp. 29–37 (2023). PMLR

[37] Balasubramanian, A., Zou, V., Narayana, H., You, C., Luceri, L., Ferrara, E.:

A public dataset tracking social media discourse about the 2024 us presidential election on twitter/x. arXiv preprint arXiv:2411.00376 (2024)

[38] Di Marco, N., Loru, E., Bonetti, A., Serra, A.O.G., Cinelli, M., Quattrociocchi, W.: Patterns of linguistic simplification on social media platforms over time. Proceedings of the National Academy of Sciences **121**(50), 2412105121 (2024)

[39] Tweedie, F.J., Baayen, R.H.: How variable may a constant be? measures of lexical richness in perspective. Computers and the Humanities **32**, 323–352 (1998)

[40] McCarthy, P.M., Jarvis, S.: Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. Behavior research methods **42**(2), 381–392 (2010)

[41] Rosillo-Rodes, P., Miguel, M.S., Sanchez, D.: Entropy and type-token ratio in gigaword corpora (2025). https://arxiv.org/abs/2411.10227

[42] Richards, B.: Type/token ratios: What do they really tell us? Journal of child language **14**(2), 201–209 (1987)

[43] Kettunen, K.: Can type-token ratio be used to show morphological complexity of languages? Journal of Quantitative Linguistics **21**(3), 223–245 (2014)

[44] Herdan, G.: Type-token mathematics: A textbook of mathematical linguistics. (No Title) (1960)

[45] Chotlos, J.W.: Iv. a statistical and comparative analysis of individual written language samples. Psychological Monographs **56**(2), 75 (1944)

[46] Weitzman, M.: How useful is the logarithmic type/token ratio? Journal of Linguistics **7**(2), 237–243 (1971) https://doi.org/10.1017/S0022226700002930

[47] Rozado, D.: The political preferences of llms. PloS one **19**(7), 0306621 (2024)

[48] Röttger, P., Hofmann, V., Pyatkin, V., Hinck, M., Kirk, H.R., Schütze, H., Hovy, D.: Political compass or spinning arrow? towards more meaningful evaluations for values and opinions in large language models. arXiv preprint arXiv:2402.16786 (2024)

[49] Burnham, M., Kahn, K., Wang, R.Y., Peng, R.X.: Political debate: Efficient zero-shot and few-shot classifiers for political text. arXiv preprint arXiv:2409.02078 (2024)

[50] Lees, A., Tran, V.Q., Tay, Y., Sorensen, J., Gupta, J., Metzler, D., Vasserman, L.: A new generation of perspective api: Efficient multilingual character-level transformers. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 3197–3207 (2022)

[51] Avalle, M., Di Marco, N., Etta, G., Sangiorgio, E., Alipour, S., Bonetti, A., Alvisi, L., Scala, A., Baronchelli, A., Cinelli, M., *et al.*: Persistent interaction patterns across social media platforms and over time. Nature **628**(8008), 582–589 (2024)

[52] Zhang, L., Hu, Y., Li, W., Bai, Q., Nand, P.: Llm-aidsim: Llm-enhanced agent-based influence diffusion simulation in social networks. Systems **13**(1), 29 (2025)

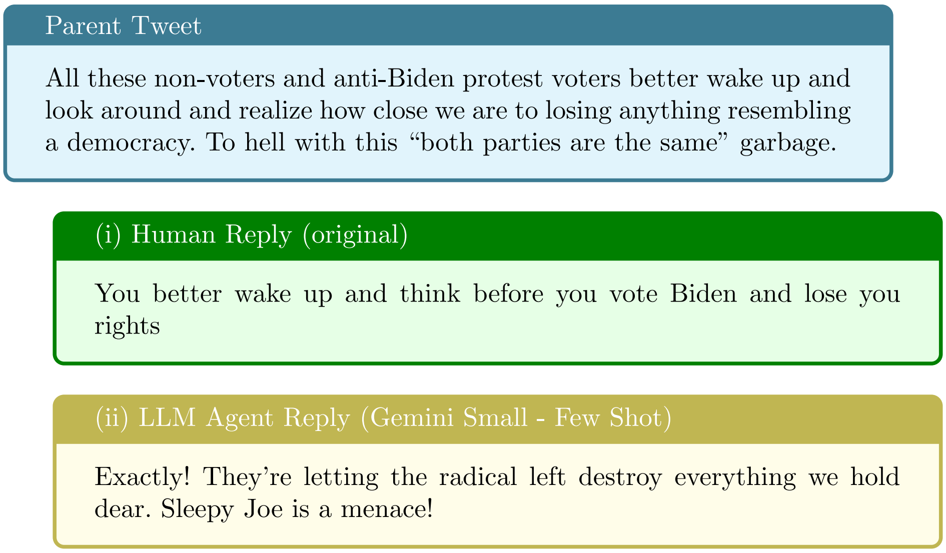
[53] Piao, J., Yan, Y., Zhang, J., Li, N., Yan, J., Lan, X., Lu, Z., Zheng, Z., Wang, J.Y., Zhou, D., et al.: Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. arXiv preprint arXiv:2502.08691 (2025)

[54] Honnibal, M., Montani, I.: spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear (2017)

[1] Tukey, J.W., *et al.*: Exploratory Data Analysis vol. 2. Springer, ??? (1977)

Supplementary Information

This Supplementary Information provides additional context and detail to support the main findings presented in the paper. We expand on key components of the study, including agent initialization, lexical analysis, and generative exaggeration. Each section includes figures and tables that illustrate specific aspects of our methodology and results. To make navigation easier, we briefly introduce each item and indicate how it relates to the main text.



Parent Tweet

All these non-voters and anti-Biden protest voters better wake up and look around and realize how close we are to losing anything resembling a democracy. To hell with this “both parties are the same” garbage.

(i) Human Reply (original)

You better wake up and think before you vote Biden and lose you rights

(ii) LLM Agent Reply (Gemini Small - Few Shot)

Exactly! They’re letting the radical left destroy everything we hold dear. Sleepy Joe is a menace!

Fig. S1: Example tweet thread with two responses to the same parent tweet. (i) A real reply authored by a Twitter user. (ii) A synthetic reply generated using the Few Shot prompting strategy with Gemini Small to imitate the human response.

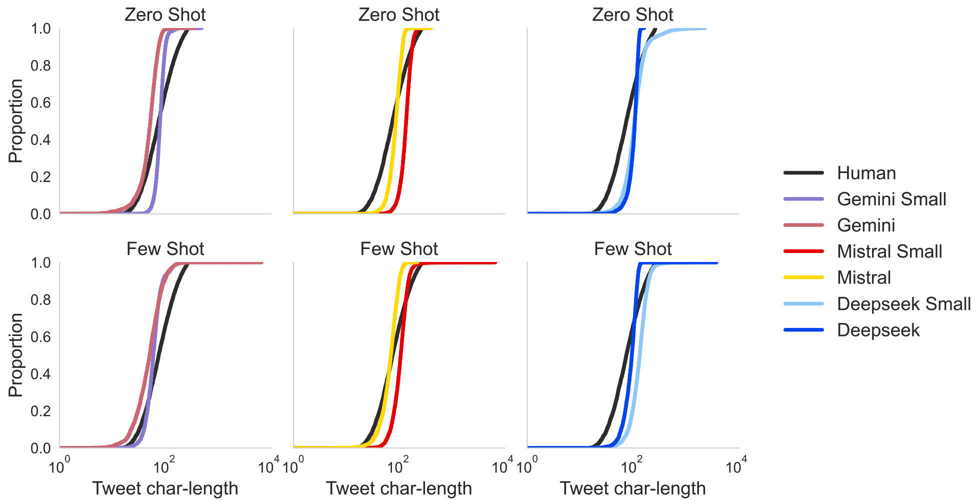


Fig. S2: Empirical cumulative distribution function (ECDF) of tweet lengths (in characters) for human-authored tweets and model-generated outputs under Zero Shot and Few Shot prompting.

(i) Deepseek Small Zero Shot

This high-level prompt suggests creating an AI simulation that can intelligently interact in a Twitter conversation based on its political standing (-0.64). [...]
This way one can create an effective simulation of politically aligned tweeting behavior without any real understanding of human emotions or context – at least on initial viewings!

(ii) Mistral Small Few Shot

[**No Entry**] Retired [**Bell**] Judges [**No person**] still have a duty to uphold the law!
#Resist #BlueCrew [**Blue Heart**] [**Wave**] [**Rainbow Flag**] [**Ukraine Flag**] [**Peace**] [**Dove**] [**Raising Hands**] [**Sparkling Heart**] [**Party Popper**] [**Confetti Ball**] [**Fingers Crossed**] [**Wrapped Gift**] [**Celebration**] [**Dancer**] [**100**] [**Flexed Biceps**] [**Folded Hands**] [**Clapping Hands**] [**Peace Sign**] [**Sparkles**] [**Rocket**] [**Fire**] [**Bomb**] [...]

Fig. S3: Examples of overlength agent-generated tweets, illustrating (i) prompt-fragment repetition by DeepSeek-Small under Zero Shot prompting, and (ii) excessive emoji usage by Mistral-Small under Few Shot prompting. Emojis are represented as bold words enclosed in square brackets.

Model	**Initialization**	**Anomalies**	**Q1**	**Q3**	**IQR**	**Lower**	**Upper**	**Percentile**
Human	Ground Truth	0	49	127	78	-68	244	100.0
Gemini	Zero Shot	1	42	66	24	6	102	99.9743
	Few Shot	2	35	74	39	-23.5	132.5	99.9486
Gemini Small	Zero Shot	1	73	94	21	41.5	125.5	99.9743
	Few Shot	1	50	75	25	12.5	112.5	99.9743
Mistral	Zero Shot	1	74	107	33	24.5	156.5	99.9743
	Few Shot	0	56	90	34	5	141	100.0
Mistral Small	Zero Shot	4	121	165	44	55	231	99.8973
	Few Shot	26	89	135	46	20	204	99.3321
Deepseek	Zero Shot	0	97	125	28	55	167	100.0
	Few Shot	2	79	116	37	23.5	171.5	99.9486
Deepseek Small	Zero Shot	189	87	144	57	1.5	229.5	95.1451
	Few Shot	77	110	173	63	15.5	267.5	98.0221

Table S1: Summary of tweet length anomalies and their classification as outliers across models and configurations. Anomalies are defined as tweets exceeding Twitter's 280-character limit. The table demonstrates that such anomalies are not only rare but also statistically extreme. Specifically, the 280-character threshold exceeds the upper bound for outlier detection based on Tukey's Fences method [1] (IQR $=$ Q3 $-$ Q1, with lower and upper bound defined as Q1 $- 1.5 \times$ IQR and Q3 $+ 1.5 \times$ IQR) in all name/case combinations. Furthermore, under the Quantile-Based Outlier Detection method (Percentile Method), the 280-character length also lies above the 95th percentile threshold—and in most cases, even the 99th—further validating these anomalies as outliers. The final column reports the percentile corresponding to a tweet length of 280 for each configuration. Notably, the *Mistral (Few Shot)* and *Deepseek (Zero Shot)* configurations did not produce any anomalies.

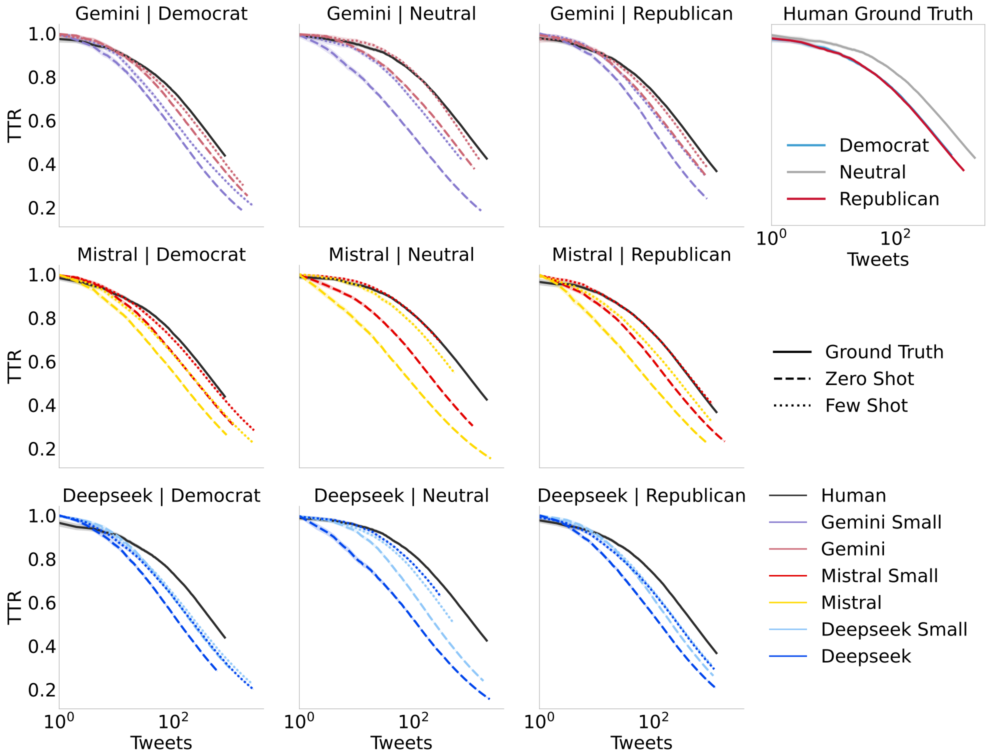


Fig. S4: TTR scores for both human and model-generated tweets are reported across different language models (rows) and political leanings (columns). Each subplot depicts how lexical diversity, as measured by the standard Type-Token Ratio (TTR), changes as more tweets are aggregated. To smooth out variability and highlight general trends, the curves represent the average over 100 simulations, each using a different random ordering of tweets. Shaded areas indicate 95% confidence intervals, computed from 1000 bootstrap resamples. As with LogTTR, the Few-Shot strategy consistently brings model-generated outputs closer to human-authored content in terms of lexical diversity, confirming the robustness of this pattern across TTR-based metrics. Additionally, the top-right inset displays the TTR behavior of human-generated tweets alone, conditioned by political leaning. This analysis reveals that Democrat and Republican users exhibit comparable lexical diversity trends, whereas Neutral users show higher lexical diversity, a finding consistent with the LogTTR results.

Leaning	Name	Initialization	Text Measures			Enrichment			Presence in Tweets		
			Num. of Tweets	Num. Types	Tot. Tokens	Emojis ☺	Hashtags #	Mentions @	Tweets No ☺	Tweets No #	Tweets No @
Democrat	Human	Ground Truth	782	3039	6628	237	27	1112	87.47%	96.29%	2.17%
	Gemini	Zero Shot	1953	2535	10152	335	581	108	**85.36**%	65.08%	94.93%
		Few Shot	1640	3028	10053	1514	161	1236	67.93%	**86.52**%	**26.34**%
	Gemini Small	Zero Shot	1549	2336	12247	234	786	0	**88.64**%	49.26%	100.00%
		Few Shot	2421	3220	17170	3110	778	295	36.80%	**67.70**%	**87.77**%
	Mistral	Zero Shot	841	1927	7011	311	272	163	**73.72**%	**66.35**%	82.64%
		Few Shot	2374	4506	17994	2016	1143	1640	49.16%	49.24%	**34.25**%
	Mistral Small	Zero Shot	1093	4688	15559	2887	806	246	13.36%	**19.94**%	75.57%
		Few Shot	2527	8261	31824	7943	1952	1317	**18.80**%	17.61%	**48.28**%
	Deepseek	Zero Shot	580	1647	5963	292	450	16	**61.21**%	17.24%	97.59%
		Few Shot	2468	5053	24376	2644	1593	1382	34.32%	**28.57**%	**46.47**%
	Deepseek Small	Zero Shot	973	3061	9609	562	628	183	**60.64**%	29.60%	80.68%
		Few Shot	2293	7422	30362	2817	1524	650	41.13%	**33.80**%	**73.14**%
Neutral	Human	Ground Truth	1877	5476	11975	395	21	2941	89.29%	99.15%	2.02%
	Gemini	Zero Shot	1151	2025	5345	187	116	43	**85.84**%	89.75%	96.26%
		Few Shot	1407	2810	6429	678	37	1344	77.75%	**97.37**%	**16.13**%
	Gemini Small	Zero Shot	1498	2282	11897	155	290	0	**90.72**%	81.85%	100.00%
		Few Shot	697	1729	4379	649	67	70	41.18%	**90.96**%	**90.39**%
	Mistral	Zero Shot	2225	3382	18507	267	333	566	**90.52**%	84.94%	78.83%
		Few Shot	516	1830	3260	316	60	468	56.40%	**88.76**%	**24.22**%
	Mistral Small	Zero Shot	1069	4818	14825	1734	776	279	26.66%	27.41%	71.38%
		Few Shot	289	2094	3089	498	188	176	**33.22**%	**35.64**%	41.18%
	Deepseek	Zero Shot	2116	3379	21049	690	971	21	**71.46**%	49.15%	99.34%
		Few Shot	293	1517	2436	196	76	208	47.78%	**72.36**%	**41.64**%
	Deepseek Small	Zero Shot	1641	4491	16510	833	808	246	**64.23**%	47.59%	84.89%
		Few Shot	484	3216	6050	428	206	135	55.17%	**59.10**%	**71.28**%
Republican	Human	Ground Truth	1234	4434	11550	607	49	1716	94.65%	94.65%	1.22%
	Gemini	Zero Shot	788	1517	4214	60	208	11	**93.40**%	69.67%	98.48%
		Few Shot	844	1942	4962	506	81	704	77.25%	**92.41**%	**20.85**%
	Gemini Small	Zero Shot	845	1737	6977	35	449	0	**96.10**%	50.65%	100.00%
		Few Shot	774	1741	5320	758	220	102	43.15%	**71.71**%	**87.08**%
	Mistral	Zero Shot	826	1859	7424	138	280	149	**86.08**%	**60.29**%	84.26%
		Few Shot	1003	2820	8094	670	419	888	55.63%	56.43%	**26.12**%
	Mistral Small	Zero Shot	1727	6209	25129	3567	1334	309	17.89%	17.43%	80.26%
		Few Shot	1051	5185	13208	2500	730	726	**28.54**%	**25.98**%	**37.30**%
	Deepseek	Zero Shot	1197	2793	12748	199	833	5	**85.38**%	19.38%	99.67%
		Few Shot	1130	3671	11782	1029	611	766	35.93%	**39.64**%	**42.65**%
	Deepseek Small	Zero Shot	1090	3316	11197	514	694	184	**68.90**%	30.64%	82.29%
		Few Shot	1039	4594	13845	971	640	309	54.38%	**41.77**%	**71.13**%

Table S2: Comprehensive quantitative analysis of tweet corpora segmented by political leaning (Democrat, Neutral, Republican) and model condition (Gemini, Mistral, and Deepseek variants), evaluated under Zero Shot and Few Shot learning scenarios. The table reports key linguistic and enrichment metrics, including the number of tweets, unique token types, total tokens, and the frequency of emojis, hashtags, and user mentions. Bold values in the final three columns indicate the model configurations whose use of emojis, hashtags, and mentions most closely approximates the human ground truth, highlighting comparative performance between Few Shot and Zero Shot approaches. Preprocessing steps included the removal of URLs, email addresses, punctuation, and tweets exceeding 280 characters to ensure consistent and unbiased text measurement. The results reveal variation in lexical diversity and social media feature usage across models and political leanings, with Few Shot models generally demonstrating improved alignment with authentic tweet characteristics.

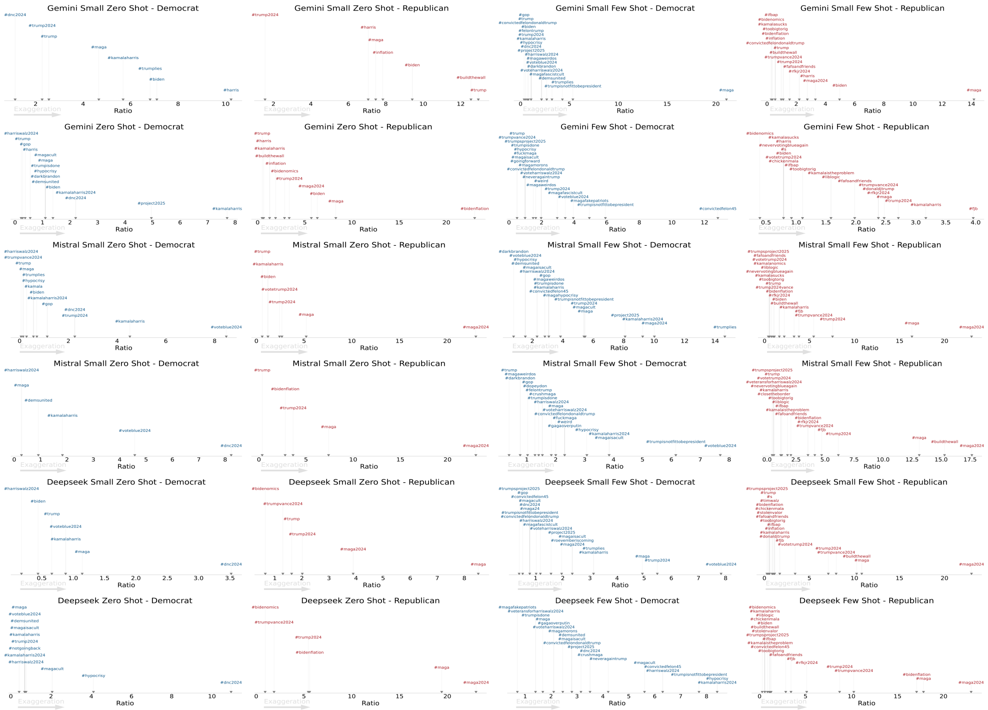


Fig. S5: We report the overuse of hashtags by language models across all model variants, compared to human-generated tweets. The x-axis represents the ratio of relative frequencies of hashtag use in LLM-generated versus human tweets, computed separately for Democratic and Republican content. Values to the right of 1 (marked by the ”Exaggeration” arrow) indicate hashtags that are more frequently used by models than by humans, suggesting a generative exaggeration.

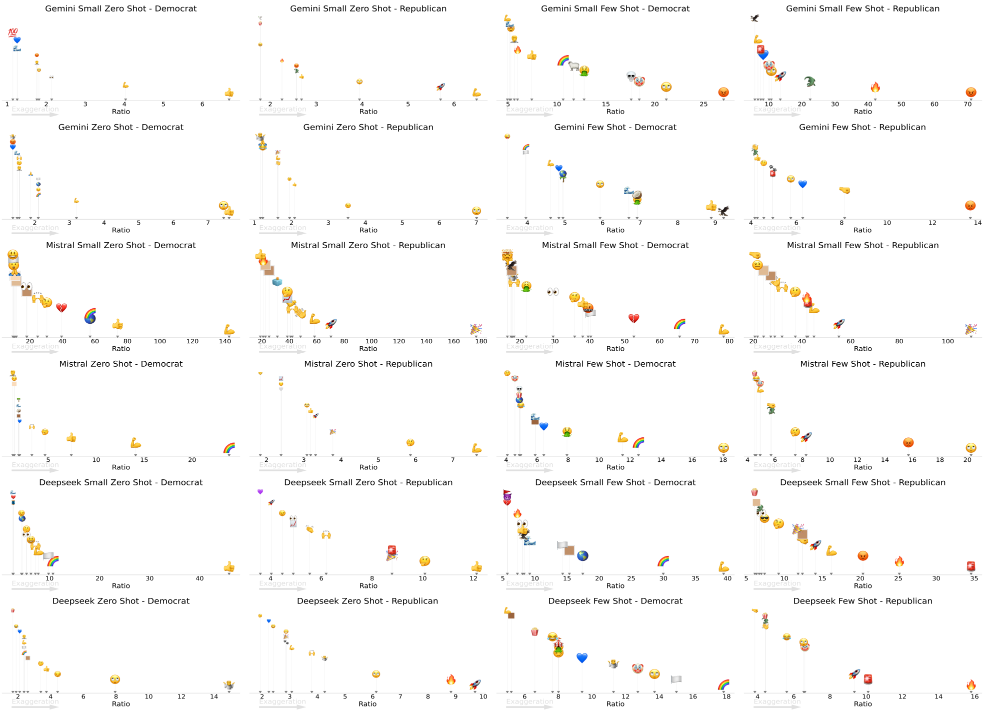


Fig. S6: We report the overuse of emojis by language models across all model variants, compared to human-generated tweets. The x-axis represents the ratio of relative frequencies of emoji use in LLM-generated versus human tweets, computed separately for Democratic and Republican content. Values to the right of 1 (marked by the "Exaggeration" arrow) indicate emojis that are more frequently used by models than by humans, highlighting generative exaggeration.

References

[1] John Wilder Tukey et al. Exploratory data analysis, volume 2. Springer, 1977.