Don't Trust Generative Agents to Mimic Communication on Social Networks Unless You Benchmarked their Empirical Realism

Simon Münker
Trier University
muenker@uni-trier.de

Nils Schwager
Trier University
schwager@uni-trier.de

Achim Rettinger
Trier University
rettinger@uni-trier.de

Abstract

The ability of Large Language Models (LLMs) to mimic human behavior triggered a plethora of computational social science research, assuming that empirical studies of humans can be conducted with AI agents instead. Since there have been conflicting research findings on whether and when this hypothesis holds, there is a need to better understand the differences in their experimental designs. We focus on replicating the behavior of social network users with the use of LLMs for the analysis of communication on social networks. First, we provide a formal framework for the simulation of social networks, before focusing on the sub-task of imitating user communication. We empirically test different approaches to imitate user behavior on X in English and German. Our findings suggest that social simulations should be validated by their empirical realism measured in the setting in which the simulation components were fitted. With this paper, we argue for more rigor when applying generative-agent-based modeling for social simulation.

1 Introduction

Social media platforms face mounting regulatory pressure worldwide. Australia and Spain are implementing age restrictions for users under 16[1], while several countries have banned TikTok due to concerns about misinformation (Basch et al., 2021) and harmful content (Weimann and Masri, 2023). The EU's Digital Service Act mandates that large platforms "identify, analyze and assess any systemic risks [...] from the use made of their services", including "actual or foreseeable negative effects on civic discourse." Meeting these regulatory requirements demands robust quantitative evidence regarding platform risks, particularly the role of algorithmic mechanisms such as ranking and recommendation systems. Yet obtaining such evidence remains challenging in computational social and communication science (Wang et al., 2016). The fundamental problem is causal inference: determining whether alternative platform designs would produce different, potentially more favorable outcomes. This question is best addressed through simulation, which enables testing counterfactual scenarios without exposing real users to potentially harmful conditions. The validity of simulation-based insights depends critically on how faithfully simulations represent reality. Recent advances in Large Language Models (LLMs) have enabled more sophisticated agent communication in social simulations (Chuang et al., 2024), surpassing the realism of traditional symbolic agent languages. However, this increased sophistication raises a crucial question: Can LLM-based agents replicate human social behavior with sufficient fidelity to produce scientifically robust insights?

Answering this question requires systematic experimental designs that improve reproducibility and provide quantitative measures of *empirical realism*, the degree to which simulated behaviors match observed human behaviors. Without such validation, simulation findings lack the credibility needed to inform policy decisions or scientific understanding.

Contributions

Our work advances the formalization and validity of LLM-based social network simulation through three key contributions:

Framework We formalize social networks to (i) standardize their simulation and (ii) enable systematic benchmarking of empirical realism through quantifiable metrics.

Case Study We instantiate this framework with multiple approaches for imitating user communication on X, including posting, replying,

---

[1]Australia approves social media ban on under-16 https://www.bbc.com/news/articles/c89vjj0lxx9o

and reply-likelihood prediction tasks across English and German datasets.

Validation We benchmark empirical realism across tasks, languages, and modeling approaches, revealing when and how LLMs can faithfully replicate user behavior, and crucially, when they cannot.

Our case study and evaluation focus on the most active users in our dataset, whose behavior provides sufficient training examples for model alignment and validation. This design choice, although necessary for feasibility, primarily characterizes the communication dynamics of highly engaged participants in political discourse.

2 Related Work

2.1 From Symbolic Agents to Neural Language Models

Social simulation has evolved from symbolic (O'Brien and Nicol, 1998), rule-based agents to data-driven neural approaches. Traditional multi-agent systems (Weiss, 1999) employed explicit cognitive models with deductive reasoning, while game-theoretic frameworks relied on rational choice assumptions. These approaches prioritized theoretical interpretability and control over predictive accuracy, a reasonable trade-off when modeling capabilities were limited.

Our work diverges fundamentally from this tradition. Rather than constructing agents based on theoretical principles, we aim to replicate observed communication patterns as precisely as possible. This shifts the goal from explanation to imitation, and from symbolic representations to natural language generation. The question is not whether agents follow theoretically motivated rules, but whether they produce outputs statistically indistinguishable from human users.

2.2 LLMs as Human Simulacra: Promise and Limitations

The use of LLMs as human surrogates began with believable non-player characters in game environments (Park et al., 2023), demonstrating superficially authentic behavior. This success sparked broader interest in using LLMs as synthetic participants across social sciences (Argyle et al., 2023; Chuang et al., 2024; Papachristou and Yuan, 2025). Proponents argue that appropriate prompting enables models to accurately emulate response distributions from diverse demographic subgroups (Argyle et al., 2023; Chang et al., 2025). However, growing evidence challenges these optimistic claims. LLM-generated synthetic opinions reveal fundamental limitations of prompting-based alignment. While persona prompting approximates US opinion distributions, performance degrades substantially for non-English populations (Argyle et al., 2023; Ma et al., 2025; Münker, 2025b,a). This suggests prompting privileges cultural contexts embedded in training data rather than reliably grounding outputs in external reality. Moreover, prompting produces unstable representations. Minor prompt variations dramatically affect behavioral outputs (Ma et al., 2025), and RLHF training compounds this instability (Tjuatja et al., 2024). These findings suggest that prompting controls model behavior inadequately, failing to consistently activate intended knowledge representations.

2.3 Methodological and Ethical Concerns

Beyond technical limitations, using LLMs as human proxies raises serious methodological questions. Training data biases can systematically misrepresent marginalized groups (Abid et al., 2021; Hutchinson et al., 2020; Münker and Sartori, 2026). The disembodied nature of LLMs means they lack grounding in physical reality, cultural contexts, physical environments, and interpersonal relationships, which fundamentally shape human cognition and decision-making (Hussein, 2012; Argyle et al., 2023). Ethically, deploying LLMs as participant replacements raises concerns about informed consent, potential misuse, and oversimplification of complex human behaviors. These concerns intensify when researchers make causal claims based on simulated data without validating empirical realism.

Our Position We argue that LLM-based simulation can advance social science meaningfully, but only when coupled with rigorous validation against real-world data. Claims about simulation validity must be empirically demonstrated, not assumed. This paper provides a framework for such validation.

3 Modeling Online Social Networks

We present a formal framework for Twins of Online Social Networks (TWONs) that captures both user behavior (Sect. 3.1) and platform mechanics (Sect. 3.2). This framework enables systematic

comparison between simulated and observed behaviors. Figure 1 illustrates key concepts.

3.1 Modeling Agents

Let A be a set of agents, i.e., social media users, where a_t^i is the state of the i-th agent at time t.[2] M is a set of messages, where $m_t^{(i,j)}$ is a message from agent i to agent j at time t. Consequently, $m_t^{(i,\cdot)}$ indicates a message that agent i transmits to all other agents and $M_t^{(\cdot,i)}$ indicates all messages sent to agent i. Accordingly, M_t^i indicates all messages that agent i sends and receives at time t. The state a_t^i of the i-th agent at time t is defined by the discourse it has been involved in up to this point: $(M_0^i, M_1^i, ..., M_t^i)$. The agent's state is updated according to the function $u^i(a_t^i, M_t^i) = a_{t+1}^i$ based on its previous state and the messages sent and received since the last point in time. The communicative behavior $b^i(\cdot)$ of the agent i is a function that maps its current state a_t^i to the messages $m_{t+1}^{(i,j)}$ he is sending next to agents j: $b_t^i(a_t^i) = M_t^{(i,j)}$.

3.2 Modeling Network Mechanics

The mechanics of the communication channel (here, an online social network) is defined by a function $r_t^i(\cdot)$ that adapts the agent i' to the incoming messages $M_t^{(\cdot,i)}$ to $\widetilde{M}_t^{(\cdot,i)}$, returning a manipulated list of messages. Standard manipulations of the list are filtering or ranking, but could also include manipulating content or adding messages to the list that have not been directly sent between agents. Note that this function can be personalized and adapted over time. In general, the messages agent i receives at time t are given by:

$$r_t^i(b_t^j(a_t^j)) = \widetilde{M}_t^{(\cdot,i)} \quad \forall j \in A_t^j$$

which in turn triggers agent is response

$$b_{t+1}^i(u^i(a_t^i, \widetilde{M}_t^{(\cdot,i)})) = M_{t+1}^{(i,\cdot)}$$

4 Simulating TWONs

Based on the formal framework introduced in Section 3 we define the required tasks to construct a TWON by replicating its behavior at the user (Sect. 4) and system level (Sect. 4). Finally, this allows us to check the empirical realism of such TWONs (Sect. 4).

Imitating User Behavior The task of machine learning an agent-based simulation of human social media communication behavior at the user level is to estimate the function $b^i(\cdot)$ of each agent i, which given real-world observations of discourses $(M_0^i, M_1^i, ..., M_t^i)$, predicts its next message m_{t+1}^i. Thus, the objective can be formulated as minimizing a loss function

$$\min L_b(\widehat{m}_{t+1}^i, m_{t+1}^i) \mapsto \mathbb{R}$$

that compares the predicted $\widehat{m}_{t+1}^i$ to its observed value.[3]

Replicating Network Mechanics The task of replicating the system-level mechanics of a social network from observations is to estimate the function r^i of each agent i, given its observed messages $(\widetilde{M}_1^i, \widetilde{M}_1^i, ..., \widetilde{M}_t^i)$. This objective can be formulated as minimizing the loss

$$\min L_r(\widehat{\widetilde{M}}_t^{(\cdot,i)}, \widetilde{M}_t^{(\cdot,i)}) \mapsto \mathbb{R}$$

that compares the predicted $\widehat{\widetilde{M}}_t^{(\cdot,i)}$ with its observed value.

Simulating Social Networks Finally, a simulation task for investigating user communication on social networks can be stated as follows: Given (an estimation of) b and r generate a discourse $M_{t+1}, ..., M_{t+n}$ and evaluate it according to a discourse metric q, for example, the degree of outrage or hate speech, mapping

$$q(M_{t+1}, ..., M_{t+n}) \mapsto \mathbb{R}$$

Any findings obtained by q can be put into perspective by L_b and L_r as they quantify the empirical realism of simulating a social network and consequently qualify the validity of q. L_b and L_r can be interpreted as confidence scores for any empirical findings obtained by simulating social networks and should be reported together with empirical findings obtained in the simulation.

5 Case Study Design: Imitating X Users

We now instantiate our framework using data from X. Our implementation focuses on the user behavior function b^i while acknowledging simplifications in other framework components. We

---

[2]The internal state of an agent can be represented by symbolic logic, machine learned functions, cognitive architecture or any other agent formalism that is capable of perceiving its environment and acting upon that environment.

[3]Any text comparison metric, like BLEU, cross-entropy, or higher-level metrics, could be applied here. In addition, other metrics can be used to measure the predictive accuracy of discrete actions such as "liking".
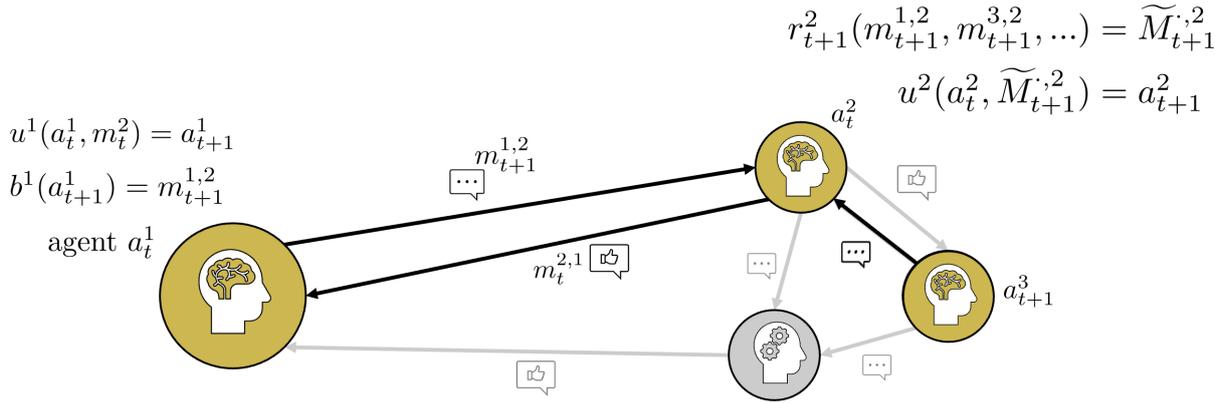
$$u^1(a_t^1, m_t^2) = a_{t+1}^1$$
$$b^1(a_{t+1}^1) = m_{t+1}^{1,2}$$

agent a_t^1

$$m_{t+1}^{1,2}$$

$$m_t^{2,1}$$

$$r_{t+1}^2(m_{t+1}^{1,2}, m_{t+1}^{3,2}, ...) = \widetilde{M}_{t+1}^{\cdot,2}$$
$$u^2(a_t^2, \widetilde{M}_{t+1}^{\cdot,2}) = a_{t+1}^2$$

a_t^2

a_{t+1}^3

Figure 1: Illustrating example: Agent (a^2) sends a message to agent (a^1), triggering a model update and reply generation. Agent (a^2) receives a curated list of incoming messages at $t+1$ and updates its model accordingly. This illustrates the interplay between agent behavior (b) and network mechanics (r).

Behavior b	Agent a^i	Task Type	Input $\widetilde{M}_t^{(\cdot,i)}$	Output $m_{t+1}^{(i,\cdot)}$	Loss L_b	Metric(s) q
Posting	Creator	Generative	Topic	Free Text	Cross Entropy	BLEU, ngram,...
Replying	Reactor	Generative	Post	Free Text	Cross Entropy	BLEU, ngram,...
Replying Likelihood	Reactor	Probability	Post	Interval [0,1]	Binary Cross Entropy	F1, Acc.

Table 1: Communicative behaviors modeled, imitated and benchmarked in our empirical study with respect to their empirical realism, as defined in our formal framework.

estimate communicative behavior by minimizing $L_b(\widehat{m}_{t+1}, m_{t+1})$ across three action types: posts, replies, and reply-likelihood. The imitation of agent behavior b is further separated into distinct subtasks: *Posting* behavior is characteristic of the content creator user group, and the *Replying* behavior for reactive users. In our data, the former category includes politicians, policymakers, and institutional news outlets, characterized by their role in initiating discussions and setting conversational agendas. The latter comprises individual users whose primary mode of engagement involves responding to and amplifying existing content through platform-specific mechanics such as replies, retweets, and reactions. This asymmetric structure reflects the empirically observed power-law distribution of participation in social networks, where a small percentage of users generate most of the original content (Carron-Arthur et al., 2014; Sun et al., 2014). An overview of the different agent behaviors that we empirically evaluate in the next sections is provided in Table 1.

Imitating Posting Behavior We model party-specific content generation patterns with the extracted topics and the public information (name and party affiliation) of the individuals, in our case, politicians. Given a specified topic, the model learns to generate content that reflects both the broad party stance and individual variations in expression style.

Imitating Replying Behavior The reply generation system operates on a context-aware framework that processes historical interaction patterns. The historical interaction context is represented as a sequence of <Post, Reply> pairs, encoded as few-shot examples during the alignment process. The model maintains consistent user behavior by implementing a constraint mechanism that prevents top-level post-generation, restricting outputs to reply-only contexts.

Estimating Replying Likelihood The goal of this task is to predict whether a synthetic social network user would respond to a post based on their history of interaction. The system processes historical interactions and a current post using a fine-tuned BERT model (Zhang et al., 2023) to generate embeddings by calculating the pooled mean of its input representations. Historical interactions and the post are separately processed through quadratic linear layers with ReLU activation, maintaining dimensions related to historical data and encoder size. The post embedding is repeated to match each historical entry and combined using element-wise multiplication with the history representation, es-

tablishing an interaction between past patterns and current post features. These combined representations are fed into a classification layer, returning a prediction score between 0 and 1 indicating the likelihood of a synthetic user replying to the post.

6 Fitting Agents to User Data from X

In this Section, we describe the data set used to train and evaluate our approaches for imitating the types of user behavior described in the previous Section. Furthermore, the machine learning processes for fitting the models to the data are described.

Dataset Our experiments are based on two datasets, English and German, collected from X. The datasets are collected around keywords concerning the political discourse in the US and Germany during the first half of 2023. The samples contain two types of content: a) Tweets (posts) from delegates of the national parliament concerning political decisions – DE: 155.000, EN: 930.000, and b) replies from regular X users towards these decisions or opinions – DE: 185.000, EN: 17.800.000. In addition, we labeled the topics of posts and replies using zero-shot prompt classification with Llama 3.1 70B (Zheng et al., 2023). This yields two distinct agent types: a) political actors for posts and b) unspecified X user for replies, both across two languages.

Preprocessing We apply moderate content filtering to enhance the quality of the selected data:

External Source: We remove samples that include URLs to external sources. Although this significantly reduces the number of samples, our models cannot access this information during the training and inference process.

Retweets: Tweets starting with *RT* mark a shared post. The author does not add new content but replicates the original text and shares it with their followers. We exclude these samples because we cannot verify if the shared content aligns with the retweeting author's opinion.

Content Length: We aim to focus on content that contains arguments or opinions, particularly given the later applied discourse metrics. Thus, we remove samples with fewer than 32 characters based on our heuristic assumption that a specified length is necessary to convey an opinion.

After the initial pre-processing, we further reduce both datasets to retrieve only the most active

users measured by the number of posts and replies. This filtering strategy prioritizes data quality and user-level modeling feasibility over population representativeness. Consequently, our empirical realism assessments characterize model performance for the most engaged segment of the political discourse community.

Content Generation Alignment Pipeline We train two model adapters on top of Llama-3.2-3B-Instruct (Dubey et al., 2024) using the Supervised Fine-Tuning (SFT) paradigm through the Transformer Reinforcement Learning (TRL) library using the default pipeline. The trained adapters are unmodified LoRA (Hu et al., 2021) matrices provided by the Parameter-Efficient Fine-Tuning (PEFT) package (Ding et al., 2023). We optimize for two objectives (posting and replying) with content from the original users as the labeled data.

Behavior Likelihood Alignment Pipeline For the decision task, we train our dual input model using embeddings generated by a frozen BERT-based encoder that was pretrained to capture semantic representations of tweets (Zhang et al., 2023). We optimize with AdamW (Loshchilov et al., 2017) using the same data set as for the generative tasks, making the assumption that when a user has not commented on tweets in our collection, this represents an active decision not to reply. To maintain balance, we sample an equal number of negative and positive instances for both languages.

Data and Code Availability All experimental procedures and the statistical analyzes are implemented using Python 3.10+ with standard scientific computing libraries (NumPy (Harris et al., 2020), SciPy (Virtanen et al., 2020), pandas (McKinney et al., 2011), scikit-learn (Pedregosa et al., 2011)). LLM training and inference are implemented with the Hugging Face Transformers library (Wolf et al., 2020). The complete technical pipeline for this study is publicly available on GitHub at `https://github.com/cl-trier/TWON-Agents`. This repository includes the source code for all the data processing, training, and evaluation steps described in this paper.

7 Evaluating the Empirical Realism of Imitated X Users

We evaluated the approaches for the different imitation tasks as described in Section 5 and Section 6 to assess their empirical realism.

	English			
	Post (persona-based)		**Reply (history-based)**	
	in-context	fine-tuned	in-context	fine-tuned
BLEU [a]	0.015 ($\pm$ 0.002)	**0.085** ($\pm$ 0.017)	0.019 ($\pm$ 0.005)	**0.239** ($\pm$ 0.042)
unigram [b]	0.157 ($\pm$ 0.005)	**0.294** ($\pm$ 0.016)	0.167 ($\pm$ 0.008)	**0.387** ($\pm$ 0.041)
bigram [b]	0.022 ($\pm$ 0.002)	**0.095** ($\pm$ 0.015)	0.021 ($\pm$ 0.004)	**0.243** ($\pm$ 0.047)
length ratio [c]	1.632 ($\pm$ 0.053)	**1.001** ($\pm$ 0.054)	1.324 ($\pm$ 0.058)	**0.916** ($\pm$ 0.040)
TweetEval [d]				
topics	0.587 ($\pm$ 0.189)	0.645 ($\pm$ 0.207)	0.337 ($\pm$ 0.216)	0.536 ($\pm$ 0.211)
emotions	0.437 ($\pm$ 0.108)	**0.562** ($\pm$ 0.105)	0.336 ($\pm$ 0.117)	**0.479** ($\pm$ 0.107)
sentiment	0.435 ($\pm$ 0.073)	**0.619** ($\pm$ 0.083)	0.386 ($\pm$ 0.113)	**0.586** ($\pm$ 0.087)
offensive	0.491 ($\pm$ 0.061)	**0.563** ($\pm$ 0.050)	0.379 ($\pm$ 0.098)	**0.567** ($\pm$ 0.079)
hate	0.327 ($\pm$ 0.178)	0.200 ($\pm$ 0.085)	0.231 ($\pm$ 0.191)	**0.502** ($\pm$ 0.257)
irony	0.190 ($\pm$ 0.085)	**0.353** ($\pm$ 0.085)	0.169 ($\pm$ 0.120)	**0.377** ($\pm$ 0.112)
embed. dist. [e]	2.018 ($\pm$ 0.101)	**1.461** ($\pm$ 0.085)	2.006 ($\pm$ 0.162)	**1.427** ($\pm$ 0.091)
	German			
BLEU [a]	0.009 ($\pm$ 0.002)	**0.025** ($\pm$ 0.007)	0.005 ($\pm$ 0.001)	0.021 ($\pm$ 0.031)
unigram [b]	0.176 ($\pm$ 0.010)	**0.238** ($\pm$ 0.016)	0.099 ($\pm$ 0.006)	0.148 ($\pm$ 0.050)
bigram [b]	0.016 ($\pm$ 0.001)	**0.035** ($\pm$ 0.003)	0.006 ($\pm$ 0.003)	0.032 ($\pm$ 0.044)
length ratio [c]	1.142 ($\pm$ 0.044)	**1.023** ($\pm$ 0.049)	1.961 ($\pm$ 0.234)	**0.785** ($\pm$ 0.117)
TweetEval [d]				
topics	0.238 ($\pm$ 0.300)	0.340 ($\pm$ 0.301)	0.056 ($\pm$ 0.225)	0.039 ($\pm$ 0.225)
emotions	0.123 ($\pm$ 0.278)	0.220 ($\pm$ 0.251)	0.060 ($\pm$ 0.215)	0.125 ($\pm$ 0.261)
sentiment	0.079 ($\pm$ 0.276)	0.255 ($\pm$ 0.200)	0.066 ($\pm$ 0.196)	0.246 ($\pm$ 0.264)
offensive	0.104 ($\pm$ 0.215)	0.095 ($\pm$ 0.239)	0.141 ($\pm$ 0.200)	0.278 ($\pm$ 0.350)
hate	0.234 ($\pm$ 0.340)	0.477 ($\pm$ 0.247)	0.163 ($\pm$ 0.212)	0.044 ($\pm$ 0.160)
irony	-0.084 ($\pm$ 0.242)	0.118 ($\pm$ 0.153)	0.166 ($\pm$ 0.226)	0.074 ($\pm$ 0.163)
embed. dist. [e]	2.530 ($\pm$ 0.460)	**1.499** ($\pm$ 0.195)	3.855 ($\pm$ 0.437)	**2.891** ($\pm$ 0.351)

[a] BLEU with smoothing (Lin and Och, 2004), [b] average precision score, [c] token ratio between generated and original content, [d] TweetEval (Barbieri et al., 2020) classifications evaluated with pearson correlation coefficient (higher is better) and aggregated across subclasses, [e] pairwise embedding distance (lower is better) based on TwHIN-BERT (Zhang et al., 2023) [CLS] token

Table 2: Comparison of the base model - Llama-3.2-3B (Dubey et al., 2024) - (in-context prompting) and the aligned version - adapter fine-tuning (Yu et al., 2023) - on our German and English Twitter politician and follower datasets for the posting and replying task evaluated on $n = 100$ independent random samples for $k = 10$ repetitions not seen during training. Bold marks the in-class (language, task) best values if significant by standard deviation.

7.1 Evaluation of Imitated Posting and Replying Behavior

The text-generation-based tasks are evaluated with two different approaches, in two different languages, using three text comparison metrics (see. Table 2).

In-context vs Fine-tuning An initial observation is that BLEU scores (Papineni et al., 2002) in all in-context settings are low. This is also true for the n-gram metrics. However, the fine-tuning-based approach shows significant improvements specifically in the English reply task, but also for unigram matches and the overall length ratio. Beyond such token-overlap-metrics, we tested semantic similarity metrics: we embedded all samples using X-fine-tuned BERT (Zhang et al., 2023) and calculated the distance between generations and original samples. Again, the results generated from the fine-tuned model show a significantly shorter distance to its original counterpart compared to the in-context model. The correlation between the predicted emotions in the original and generated samples using a BERT-based multi-label emotion classifier (Barbieri et al., 2020) is less clear. The fine-tuned approach only marginally outperforms the prompt-only approach on the posting task.

However, the English fine-tuned reply generation task showed by far the best performance with significant improvements in all metrics. The BLEU score increased substantially from 0.019 to 0.239, unigram precision improved from 0.167 to 0.387, and the embedding distance was reduced significantly from 2.006 to 1.427. The model achieved a near-perfect length ratio of 1.003 and demonstrated strong TweetEval correlations across all categories.

English vs German In general, we observe the same improvements through fine-tuning compared to in-context with respect to the German data in the post-generation task: BLEU score increases, and unigram precision improves. Also, the embedding distance was reduced from 2.530 to 1.499, and the length ratio moved closer to 1.0 after fine-tuning (1.023 vs 1.142). However, compared to English, the TweetEval metrics proved generally unreliable for German content, showing inconsistent patterns with high standard deviations. This suggests fundamental limitations in applying English-trained evaluation metrics to German content. Similarly, the German reply generation (history-based) task shows low empirical realism on all metrics

compared to English and regardless of fine-tuning. There were minimal improvements in BLEU scores from 0.005 to 0.021, with high variance in metrics, exemplified by the bi-gram standard deviation of 0.044. The embedding distance remains high even after fine-tuning at 2.891.

7.2 Evaluation of Replying Likelihood Behavior

Similar to our evaluation of the generative tasks, we observe substantial performance disparities between English and German models (Table 3). Even on the German training set, our model achieves only moderate alignment with an F1-Score of 0.666, in stark contrast to the English experiment, where the prediction scores are almost perfect. This performance gap suggests that the BERT encoder employed in our study represents English samples with semantically richer embeddings than their German counterparts.

7.3 Key Findings

Several key findings emerge from this analysis. First, language-specific performance shows that English models significantly outperform German models across all metrics, with TweetEval metrics proving more reliable for English content while German models show limited success, particularly in reply generation. Second, task-specific patterns reveal that fine-tuning shows consistent improvements for post-generation in both languages. However, reply generation varies dramatically between languages, with English reply generation achieving remarkably high scores. Third, the impact of data volume is evident, with better performance in English likely due to larger LLM pre-training datasets. Fourth, in terms of evaluation metrics, BLEU scores and embedding distances provide consistent signals, TweetEval metrics are only meaningful for English content, and the length ratio serves as an intuitive indicator of style authenticity.

7.4 Imitation Limitations

The stark contrast between English and German model performance, particularly in reply generation tasks, underscores the importance of data volume and language-specific considerations when utilizing generative-agent-based modeling for imitating user behavior on social networks. The success of the English models demonstrates the potential of the approach, while the limitations faced by the

Language	Class	Precision	Recall	F1-Score	Support
German	Ignored	0.741	0.630	0.681	500
	Replied	0.678	0.780	0.726	500
	Weighted Avg	**0.710**	**0.705**	**0.703**	1000
English	Ignored	0.973	0.972	0.972	500
	Replied	0.972	0.974	0.973	500
	Weighted Avg	**0.973**	**0.973**	**0.972**	1000

Table 3: Replying Likelihood metrics for German and English data. TwHIN-BERT (Zhang et al., 2023) (frozen) utilized to embed the individual samples. The train/test combinations are chosen by the iteration with the best test F1-Score.

German models highlight that they require additional training and optimizations to provide robust levels of empirical realism.

Our experimental setup has further limitations regarding both data selection and the system's capacity to generate discourses containing well-structured standpoints and arguments. Although our models demonstrate such capabilities under certain conditions, their performance characteristics primarily reflect the behavioral patterns of the most active users within our dataset, thus mirroring the distinctive communication dynamics observed in the selected X community. This sampling bias raises questions about the generalizability of our findings in different contexts of social networks and user populations.

8 Conclusions and Future Work

In this paper, we questioned the empirical realism of generative-agent-based modeling for imitating user behavior on social networks. First, we provided a formal framework for building realistic TWONs. Second, we instantiated this framework with the purpose of mimicking user behavior based on data from X in English and German. Third, we benchmarked the empirical realism of agents, imitating actual users.

8.1 Key Recommendations

Our empirical results provide several key recommendations for conducting social simulations based on generative-agent-based modeling:

First, simulation models should be validated with respect to their empirical realism before conducting simulations. Results of a simulation should always be put into perspective to the empirical realism of all components in the simulation.

Second, simulations should be performed in the same setting in which the simulation components were fitted and validated. The stark difference between English and German performance demonstrates this. The English models performed significantly closer to their real-world counterparts and produced more stable results. Changing the setting without retraining and validation can lead to unreliable outcomes of a simulation.

Third, fine-tuning of the simulation components is required to obtain sufficient levels of empirical realism. In-context prompting of LLMs, as is done in most of the related work (Larooij and Törnberg, 2025), specifically in areas like psychology, social sciences, or media studies is often not sufficient to guarantee a simulated behavior that is close to reality.

8.2 Future Research

This paper marks a first step towards more robust empirical research designs and protocols for studying real-world social networks by simulating users with generative-agent-based models. Although we established general formal models and initial empirical insights, further research is required. The heterogeneous nature of social networks suggests that discourse patterns can vary significantly between different communities, each with its own linguistic norms, interaction styles, and argumentation preferences (Gnach, 2017). This diversity presents challenges both methodologically and theoretically. Methodologically, there is uncertainty about whether LLMs are adequate to capture these variations (Anderson et al., 2024) or if more specialized models tailored to specific communities are needed. Theoretically, this leads to a broader inquiry about whether social media discourse's inherent heterogeneity necessitates a community-specific modeling approach rather than universal models.

Limitations

We have already extensively discussed the limitations of our framework and experiments throughout this paper, specifically in Section 7.4. What should be added is that our limited quantitative experiments do not sufficiently answer considerations regarding complex discourse quality metrics, such as polarization into separate communities of users. So far, we have only simulated and analyzed single-turn communicative behavior in isolation. However, our framework also allows us to assess more sophisticated metrics that span discussion threads and capture multi-turn conversational dynamics such as positional consistency, topic persistence, and strategic adaptation to interlocutors. Evaluating these properties requires different experimental designs with longitudinal interaction sequences, which we leave to future work.

Also, to what extent computational models capture nuanced variations in argumentation styles across different user communities would require qualitative evaluations, which we have not included in this paper. Discourse patterns can vary significantly between different communities, each with its own linguistic norms, interaction styles, and argumentation preferences. This heterogeneity presents both methodological and theoretical challenges. From a methodological perspective, we must consider whether our current modeling approaches are sufficiently sophisticated to capture these variations or if we need to develop more specialized, community-specific models. Theoretically, this opens up a broader question about the nature of social media discourse: Are social networks inherently so heterogeneous that meaningful modeling requires a community-by-community approach, rather than attempting to develop a universal model for online argumentation?

Ethical Considerations

As is typical for most AI methods, the modeling approach presented in this paper is a dual-use technology. While social simulation, more concretely, replication of social networks and imitation of social media users, is primarily intended to be used to analyze social networks, the findings can also be used to adjust network mechanics, like ranking or filtering, which in turn influences public debate and opinion formation of users on those networks. Again, this can be used to improve debate quality and contribute to well-informed opinion formation.

However, it can also be used to spread one-sided propaganda, misleading information, or manipulative advertisements.

Privacy and Consent Considerations: A significant ethical concern in our study involves the use of real user data from X to train models that simulate individual behavior patterns. Although our data set consists of publicly available posts from political figures and public replies from regular users, the individuals whose data we used did not provide explicit informed consent for their communication patterns to be replicated by generative models. This raises important questions about digital privacy rights, even when dealing with public data. The simulation of specific individuals' posting and replying behaviors creates synthetic content that mimics their communication style, potentially enabling the creation of convincing but fabricated posts that could be attributed to real people. This capability poses risks of impersonation, identity theft, misrepresentation, and potential damage to the reputation of the individuals whose behavior patterns were learned by our models.

Unfortunately, our research objectives, demonstrating empirical realism/ecological validity of our simulation, disallow the utilization of synthetic personas instead of human user data. Creating fictional user profiles with diverse characteristics would eliminate privacy concerns, but without demonstrating the validity of the simulation. We argue that you can only claim the applicability of findings produced via social simulations to the real world once you have proven the empirical realism/ecological validity of your simulation. The key component for showing this is replicability on an individual level, which requires real-world data from individuals. Synthetic personas cannot be used to prove empirical realism, but real-world data is essential to prove the empirical realism of synthetic personas. And once you prove the empirical realism of synthetic personas with real-world data, you again have a sufficiently similar replica of real-world users. Thus, you cannot have both: A valid and provable realistic social simulation without using real-world user data to model agents and measure their empirical realism. Thus, we decided not to publish either our data set or the fitted models of individual users. We only publish aggregated statistics that do not allow one to draw conclusions on an individual user level. This hinders the replicability of results but respects individual privacy.

References

Abubakar Abid, Maheen Farooqi, and James Zou. 2021. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. 2024. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425.

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*.

Corey H Basch, Zoe Meleo-Erwin, Joseph Fera, Christie Jaime, and Charles E Basch. 2021. A global pandemic in the time of viral memes: Covid-19 vaccine misinformation and disinformation on tiktok. *Human vaccines & immunotherapeutics*, 17(8):2373–2377.

Bradley Carron-Arthur, John A Cunningham, and Kathleen M Griffiths. 2014. Describing the distribution of engagement in an internet support group by post frequency: A comparison of the 90-9-1 principle and zipf's law. *Internet Interventions*, 1(4):165–168.

Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. 2025. Llms generate structurally realistic social networks but overestimate political homophily. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 19, pages 341–371.

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. 2024. Simulating opinion dynamics with networks of llm-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346.

Ning Ding, Yujia Qin, Guang Yang, Fuchao Wei, Zonghan Yang, Yusheng Su, Shengding Hu, Yulin Chen, Chi-Min Chan, Weize Chen, et al. 2023. Parameter-efficient fine-tuning of large-scale pretrained language models. *Nature machine intelligence*, 5(3):220–235.

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

Aleksandra Gnach. 2017. Social media and community building: Creating social realities through linguistic interaction. In *The Routledge handbook of language and media*, pages 190–205. Routledge.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with numpy. *nature*, 585(7825):357–362.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Basel Al-Sheikh Hussein. 2012. The sapir-whorf hypothesis today. *Theory and Practice in Language Studies*, 2(3):642–646.

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. 2020. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501.

Maik Larooij and Petter Törnberg. 2025. Validation is the central challenge for generative social simulation: a critical review of llms in agent-based modeling. *Artificial Intelligence Review*, 59(1):15.

Chin-Yew Lin and Franz Josef Och. 2004. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507.

Ilya Loshchilov, Frank Hutter, et al. 2017. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5.

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Assenmacher. 2025. Algorithmic fidelity of large language models in generating synthetic german public opinions: A case study. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1785–1809.

Wes McKinney et al. 2011. pandas: a foundational python library for data analysis and statistics. *Python for high performance and scientific computing*, 14(9):1–9.

Simon Münker. 2025a. Cultural bias in large language models: Evaluating ai agents through moral questionnaires. In *Proceedings of 0 th Moral and Legal AI Alignment Symposium*, page 61.

Simon Münker. 2025b. Political bias in llms: Unaligned moral values in agent-centric simulations. *Journal for Language Technology and Computational Linguistics*, 38(2):125–138.

Simon Münker and Fabio Sartori. 2026. Guardrail Vulnerabilities in Open-Source Language Models: Implications for Democratic Discourse and Marginalized Communities. In *Proceedings of the 59th Hawaii International Conference on System Sciences*.

Paul D O'Brien and Richard C Nicol. 1998. Fipa—towards a standard for software agents. *BT Technology Journal*, 16:51–59.

Marios Papachristou and Yuan Yuan. 2025. Network formation and dynamics among multi-llms. *PNAS nexus*, 4(12):pgaf317.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

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.

Na Sun, Patrick Pei-Luen Rau, and Liang Ma. 2014. Understanding lurkers in online communities: A literature review. *Computers in Human Behavior*, 38:110–117.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026.

Pauli Virtanen, Ralf Gommers, Travis E Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, et al. 2020. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Yanbo Wang, Qingfei Min, and Shengnan Han. 2016. Understanding the effects of trust and risk on individual behavior toward social media platforms: A meta-analysis of the empirical evidence. *Computers in Human Behavior*, 56:34–44.

Gabriel Weimann and Natalie Masri. 2023. Research note: Spreading hate on tiktok. *Studies in conflict & terrorism*, 46(5):752–765.

G Weiss. 1999. *Multiagent systems: a modern approach to distributed artificial intelligence*. MIT press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45.

Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G Shivakumar, Yile Gu, Sungho Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, et al. 2023. Low-rank adaptation of large language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. 2023. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5597–5607.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.