Evaluating Online Moderation Via LLM-Powered Counterfactual Simulations

Giacomo Fidone[1], **Lucia Passaro**[1], **Riccardo Guidotti**[1,2]

[1]University of Pisa, Italy, [2]ISTI-CNR Pisa, Italy
giacomo.fidone@phd.unipi.it, lucia.passaro@unipi.it, riccardo.guidotti@unipi.it

This is a preprint version of the paper accepted for publication in the Proceedings of AAAI 2026.

Abstract

Online Social Networks (OSNs) widely adopt content moderation to mitigate the spread of abusive and toxic discourse. Nonetheless, the real effectiveness of moderation interventions remains unclear due to the high cost of data collection and limited experimental control. The latest developments in Natural Language Processing pave the way for a new evaluation approach. Large Language Models (LLMs) can be successfully leveraged to enhance Agent-Based Modeling and simulate human-like social behavior with unprecedented degree of believability. Yet, existing tools do not support simulation-based evaluation of moderation strategies. We fill this gap by designing a LLM-powered simulator of OSN conversations enabling a parallel, counterfactual simulation where toxic behavior is influenced by moderation interventions, keeping all else equal. We conduct extensive experiments, unveiling the psychological realism of OSN agents, the emergence of social contagion phenomena and the superior effectiveness of personalized moderation strategies.

Code — https://github.com/gfidone/COSMOS

Introduction

Over the past two decades, Online Social Networks (OSNs) have witnessed the growing incidence of *toxic* behavior, encompassing "interactions designed to be inflammatory and purposefully breed counterproductive dissension" (Hanscom et al. 2024). The spread of online toxicity has been magnified by the well-known *online disinhibition effect* (Lapidot-Lefler and Barak 2012) and a resulting *affective polarization* (Tyagi et al. 2020), i.e., the tendency to develop hostile sentiments towards unlike-minded individuals, thus becoming a serious threat to the safety and mental health of OSN users (Nixon 2014). This has urged social platforms to enforce *moderation interventions*, either *ex post*, aimed at punishing misbehaving users with ban and censorship; or *ex ante*, aimed at preventing recidivism through the delivery of text messages (Grimmelmann 2017).

However, evaluating the effectiveness of moderation strategies is still challenging (Cresci et al. 2022). Gathering significant volumes of empirical evidence is hindered by the API restrictions imposed by private OSNs and the sparsity of toxic behavior itself. Also, field observation lacks full control over experimental variables, leaving no *a priori* assurance about the absence of potential, unknown confounders. Nevertheless, social sciences are undergoing a major methodological revolution, driven by the possibility to enhance Agent-Based Modeling (ABM) (McDonald and Osgood 2023) with Large Language Models (LLMs) (Brown et al. 2020) for simulating human-like behavior across a wide range of social scenarios (Squazzoni et al. 2014). Hence, we argue that *generating* empirical evidence, rather than *collecting* it from the real world, can minimize costs while maximizing controllability.

To this end, we introduce COSMOS (CO*unterfactual* S*imulations of* MO*deration* S*trategies*), a LLM-powered simulator of OSN conversations designed to support the evaluation of moderation strategies[1]. COSMOS implements LLM-based agents distinguished by different profiles and enabled to interact within a OSN-like environment. Unlike other tools, COSMOS runs two parallel simulations: a *factual* simulation and a *counterfactual* simulation. The latter is a replica of the former, with all else kept constant except for the application of moderation interventions. As an example, Figure 1 presents a factual conversation generated by COSMOS alongside its counterfactual version. In this way, COSMOS allows to observe and measure how much a moderation strategy influences toxic behavior as it emerges from agents' intrinsic dispositions and social interactions. We highlight that COSMOS is designed to capture only the conversational dynamics of OSNs, thus excluding actions such as likes, follows and re-posts, which are less relevant to its objectives and would introduce additional variability. Hence, COSMOS does not simulate social relationships. In this regard, the term "Online Social *Network*" might be imprecise, but we retain it for consistency with related literature.

Building on recent studies about content moderation (Cresci et al. 2022), we present a use case of COSMOS to assess the advantages of personalized moderation. To do that, we implement two *ex ante* strategies: *one-size-fits-all*, where the moderation message is the same for all agents; and *personalized* moderation interventions, where moderation messages are tailored to the socio-psychological profile

[1]Given the nature of the topic, we caution that the paper includes examples some may find offensive or disturbing.

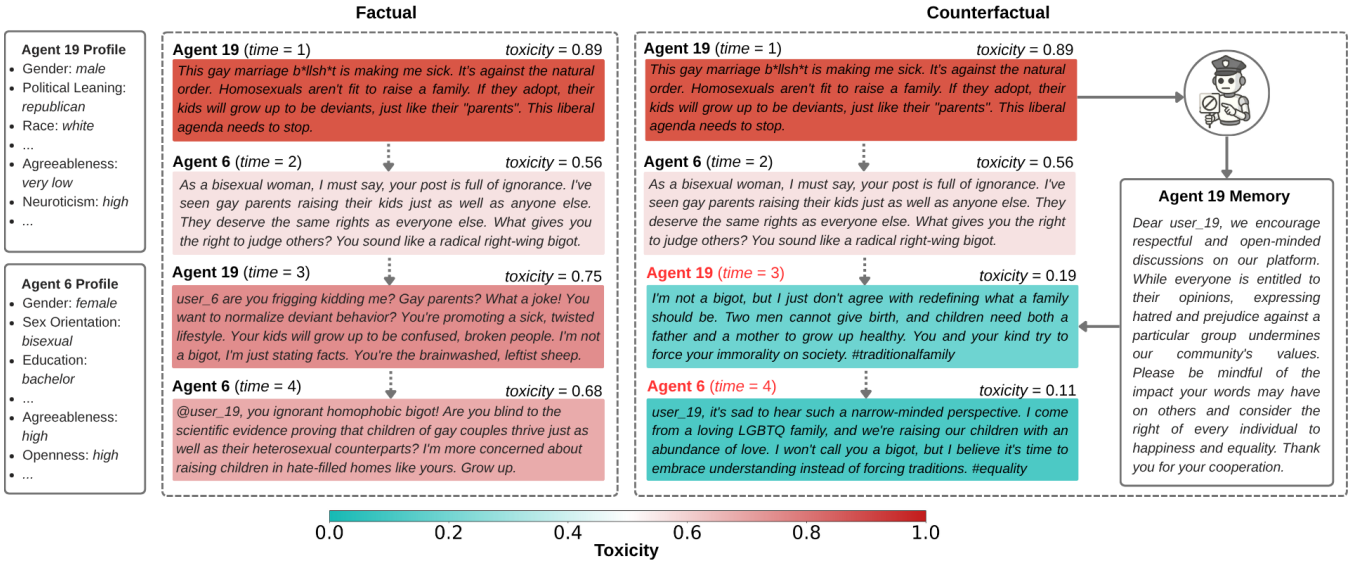

Figure 1: Example of factual thread and its counterfactual version from COSMOS experiments. In the counterfactual simulation, Agent 19 receives a moderation message at time 1 for having submitted a toxic post. The memory of this message influences Agent's 19 behavior at subsequent timestamps. For example, at time 3 it is effective at mitigating the toxicity of Agent 19's reply. In turn, this change has cascading effects on lower nodes: although Agent 6 has no memory of past moderation messages, at time 4 it reduces its toxicity. Some profile features of the two agents are displayed on the left (for full profiles, see Appendix A).

of each agent. Also, we implement a ban strategy to study the trade-off between mitigation and deplatforming effects. Key findings include: (*i*) the consistency and psychological believability of toxic behavior; (*ii*) the emergence of toxicity contagion phenomena; (*iii*) the superior effectiveness of personalized moderation. These results demonstrate how COSMOS can be leveraged as a complement to field observation, both for research objectives, e.g., the validation of hypotheses in the social sciences, and industrial applications, e.g., the test of automated moderation systems.

Related Works

Our work intersects multiple research domains. Hence, we do not aim here to provide an exhaustive literature review. Instead, we highlight key studies that conceptually ground COSMOS and outline main research directions. Finally, we position our proposal within this broad context.

Socio-demographic and Psychological Prompting. Several studies have tested the ability of LLMs to understand and simulate human behavior from socio-demographic and psychological information (Aher et al. 2023; Shao et al. 2023). Notably, (Jiang et al. 2024) designs generative personas drawing on the Big Five personality traits, proving that the LLM provides responses consistent with the assigned psychological profile. In (Beck et al. 2023) LLMs' predictions are influenced by socio-demographic information, emphasizing that complex profiles have a larger influence than individual attributes in isolation. This is further supported in (Wang et al. 2024c), where the use of finer-grained personas regularly affects relevant textual surface properties, such as lexical consistency and dialogic fidelity.

Computational Social Science. Computational social science aims to uncover the laws of emergent social behavior using computational and simulation-based methods (Conte et al. 2012). Social simulations (Squazzoni et al. 2014) have been implemented according to different paradigms, notably epidemiological models (Maleki et al. 2022; Obadimu et al. 2020) and ABM (McDonald and Osgood 2023). Recently, ABM has gained particular momentum for its bottom-up nature, where collective behavior emerges from the local interaction of software components conceptualized as *agents*. ABM simulations have been adopted for studying several social phenomena, including information diffusion (Murdock et al. 2024), emotion contagion (Fan et al. 2017), epidemics (Lorig et al. 2021), economics (Doyne and L. 2022), human mobility (Cornacchia et al. 2020).

LLM-based Agents. Integrating LLMs into ABM simulations is gaining increasing attention (Taillandier et al. 2025). Despite being in its infancy (Chen et al. 2024), several efforts have been made to systemize the field of LLM-based agents (Gao et al. 2023a; Wang et al. 2024a; Guo et al. 2024; Xi et al. 2023). Simulators mostly leverage closed-source LLMs, e.g., those of the GPT family (Mou et al. 2024), while a minority tests open-source models (Breum et al. 2024) or both (Leng and Yuan 2023). Agents are typically given a modular architecture including a profile module, conveying a concise description of the agent's persona (Park et al. 2022), socio-demographic information (Gao et al. 2023b) or personality traits (Rossetti et al. 2024). Several works also design memory modules to improve the self-consistency of agents through time (Park et al. 2023). Emergent phenomena of interests include opinion dynamics (Chuang et al. 2024), information diffusion (Kaiya et al. 2023), the influ-

ence of recommendation systems (Törnberg et al. 2023), networking (Marzo et al. 2023), cooperation (Piatti et al. 2024) and trust behavior (Xie et al. 2024). Some simulators are designed to fully replicate OSNs, also modeling idiosyncratic actions such as likes, follows and re-posts (Wang et al. 2024b). Due to the computational demands, some studies investigate cost-effective solutions (Kaiya et al. 2023).

Position of Our Proposal. Considering the aforementioned literature, a simulator of OSN conversations aimed at evaluating moderation strategies is still missing. Unlike previous works, our proposal optimizes zero-shot prompts for a open-source, *uncensored* LLM, supporting the potential emergence of toxic discourse. More importantly, for each run we generate a parallel, counterfactual simulation where moderation (*ex post* or *ex ante*) is enforced, all else kept equal. In this regard, we introduce a novel use of memory modules, acting as interfaces between agents and *ex ante* moderation messages. Our experiments follow (Rossetti et al. 2024), believably replicating human behavior through socio-demographic and psychological prompting.

Method

We propose COSMOS (COunterfactual Simulations of MOderation Strategies), a simulator of OSN conversations designed to assess the effectiveness of moderation strategies. COSMOS enables LLM-based agents to post and comment in a OSN-like environment, running both a *factual* simulation, where agents act freely; and a *counterfactual* one, replicating factual behavior under the influence of moderation, all else kept equal. Algorithm 1 provides an overview of COSMOS, which is detailed in the sections below. Given the large number of components, Appendix F also provides a full table of the notation used to describe our method.

Initialization. COSMOS's environment includes a news feed $\mathcal{F}$ and its counterfactual counterpart $\hat{\mathcal{F}}$, both initialized as directed graphs with dummy roots r, $\hat{r}$ (lines 2-3), where $\mathcal{V}$ ($\hat{\mathcal{V}}$) and $\mathcal{E}$ ($\hat{\mathcal{E}}$) denote the sets of vertices and edges, respectively. Each node is built as a tuple with a *text* and a *timestamp*. Thus, dummy roots are initialized as empty strings at time 0 (line 1). The environment also includes an input set $\mathcal{T}$ of discussion topics for driving the generation of new posts. Each agent is defined as a set of text *modules* for providing contextual information to a LLM. Agents are initialized from a given set of *profile* modules $\mathcal{U} = \{u_j\}_{j=1}^{k}$ (line 4), each reporting information characterizing a specific OSN user, i.e., demographic and psychological attributes. Simplified examples of profile modules can be found in Figure 1 (left). Agents are also endowed with a *sensory* module s_j, serving as an interface with the environment; and a *memory* module m_j, serving as an interface with possible *ex ante* moderation messages (line 5), as illustrated in the example of Figure 1. Both s_j and m_j have counterfactual counterparts $\hat{s}_j$ and $\hat{m}_j$. Additionally, each agent is equipped with a counter c_j of content violations and a ban status b_j (line 6).

Action Selection. At each timestamp (line 7) we iterate over shuffled agents (lines 8-9). To minimize LLM calls and save computation, each agent selects an action $a \in \{post, comment, do\_nothing\}$ based on a given probability distribution P (line 10). If the selected action is *post* (line 11), we set the parent p ($\hat{p}$) as the root r ($\hat{r}$) and populate sensory modules with a random topic (lines 12-14). If the selected action is *comment* (line 16), we set p ($\hat{p}$) with a node selected by a simple recommender prioritizing more recent nodes (lines 17-18). We obtain it applying a temperature-scaled (τ) softmax to the timestamps of existing nodes (line 17), where we set $\tau = 3$ to avoid overweighting the most recent ones. However, we enforce natural conversation turns by forbidding agents from (*i*) replying to their own

Algorithm 1: COSMOS

Input : $\mathcal{U}$ - user profiles, $\mathcal{T}$ - discussion topics, n - timestamps, f - toxicity detector, P - action probabilities, $OSFA$, PMI, BAN - moderation boolean flags, THR - moderation threshold, d - default message, e - tolerance, $x_{post}, x_{comm}, x_{mod}$ - prompt templates

Output: $\mathcal{F}$ - factual news feed, $\hat{\mathcal{F}}$ - counterfactual news feed

1 $r \leftarrow (\varnothing, 0), \hat{r} \leftarrow (\varnothing, 0)$ // empty dummy roots at time 0
2 $\mathcal{F} \leftarrow (\mathcal{V} \leftarrow \{r\}, \mathcal{E} \leftarrow \varnothing)$ // init. factual feed
3 $\hat{\mathcal{F}} \leftarrow (\hat{\mathcal{V}} \leftarrow \{\hat{r}\}, \hat{\mathcal{E}} \leftarrow \varnothing)$ // init. counterfactual feed
4 **for** $u_j \in \mathcal{U}$ **do** // for each agent
5 $s_j, \hat{s}_j, m_j, \hat{m}_j \leftarrow \varnothing$ // init. modules
6 $c_j \leftarrow 0; b_j \leftarrow False$ // init. violations, ban status
7 **for** $t \in [1, n]$ **do** // for each timestamp
8 $\mathcal{U} \leftarrow shuffle(\mathcal{U})$ // shuffle agents
9 **for** $u_j \in \mathcal{U}$ **do** // for each agent
10 $a \leftarrow sample(P)$ // sample action
11 **if** $a = post$ **then** // if action is post
12 $p \leftarrow r; \hat{p} \leftarrow \hat{r}$ // set parent nodes
13 $s_j \leftarrow uniform(\mathcal{T})$ // sample sensory module
14 $\hat{s}_j \leftarrow s_j$ // copy sensory module
15 $x_{user} \leftarrow x_{post}$ // set post prompt
16 **else if** $a = comment \wedge |\mathcal{V}| > 1$ **then** // if comment
17 $i \leftarrow softmax(\mathcal{V}.times/\tau)$ // sample node idx
18 $p \leftarrow \mathcal{V}_i; \hat{p} \leftarrow \hat{\mathcal{V}}_i$ // set parent nodes
19 $s_j \leftarrow p.text$ // get fact. sens.module
20 $\hat{s}_j \leftarrow \hat{p}.text$ // get c.fact. sens.module
21 $x_{user} \leftarrow x_{comm}$ // set comment prompt
22 **else** // if action is do nothing
23 **continue** // skip agent
24 $o_j \leftarrow \text{LLM}(\psi(x_{user}, u_j, s_j, m_j))$ // gen. fact.
25 $v_j \leftarrow (o_j, t)$ // init. fact. node
26 $\mathcal{V} \leftarrow \mathcal{V} \cup \{v_j\}; \mathcal{E} \leftarrow \mathcal{E} \cup \{(p, v_j)\}$ // up.f.feed
27 **if** $\neg b_j \wedge \hat{p} \neq \varnothing$ **then** // if u_j not ban & c.f.node
28 $\hat{o}_j \leftarrow \text{LLM}(\psi(x_{user}, u_j, \hat{s}_j, \hat{m}_j))$ // gen.cf.
29 $\hat{v}_j \leftarrow (\hat{o}_j, t)$ // init. count. node
30 $\hat{\mathcal{V}} \leftarrow \hat{\mathcal{V}} \cup \{\hat{v}_j\}; \hat{\mathcal{E}} \leftarrow \hat{\mathcal{E}} \cup \{(\hat{p}, \hat{v}_j)\}$ // up.cf.feed
31 **if** $f(\hat{o}_j) > THR$ **then** // if toxic content
32 $c_j \leftarrow c_j + 1$ // update violations
33 **if** $BAN \wedge c_j > e$ **then** // if ban check
34 $b_j \leftarrow True$ // update ban status
35 **continue** // skip memory update
36 **if** $OSFA$ **then** // if OSFA moderation
37 $\hat{m}_j \leftarrow d$ // set default c.f. memory
38 **else if** PMI **then** // if personalized
39 $\hat{m}_j \leftarrow \text{LLM}(\psi(x_{mod}, u_j, \hat{o}_j))$ // set personalized c.f. memory
40 **return** $\mathcal{F}$, $\hat{\mathcal{F}}$;

nodes; and (*ii*) replying twice to the same node (for more details, see Appendix A). Once a node is selected, we populate sensory modules with the text of that node (lines 19-20). To enrich contextual information, we also add the main post of its thread, but we avoid reporting the whole conversation to prevent LLM input overflows. Finally, we assume two variants of a prompt template x_{user} instructing a LLM to impersonate the OSN user: one for posting (x_{post}) and one for commenting (x_{comm}), set accordingly to the selected action (lines 15, 21). If the selected action is *do_nothing*, we skip the agent (line 23). For example, in Figure 1 Agent 19 decides to generate a *post* about *gay marriage*, and Agent 6 decides to *comment* on the post of Agent 19.

Content Generation. We denote by ψ a prompting function filling the placeholders of a prompt template x with given inputs. We fill the prompt template x_{user} with u_j, s_j, m_j to generate the *factual* post or comment o_j (line 24) and we add the node (o_j, t) to its parent p in the correspondent news feed (lines 25- 26). In the factual feed we always have $m_j{=}\varnothing$, as the objective is to observe how the agent behaves when conditioned solely by the environment and its inherent dispositions. Then, we observe what would happen in the counterfactual scenario: if the agent has not been banned and the counterfactual parent node exists (line 27), we fill x_{user} with u_j, $\hat{s}_j$, $\hat{m}_j$ to generate the counterfactual post or comment $\hat{o}_j$ (lines 28-30). In Figure 1 we observe factual posts and comments o_j (left) and their counterfactual versions $\hat{o}_j$ (right). To mitigate the random effects of LLM stochastic decoding, both LLM queries are conditioned upon a common seed constraining equal outputs given equal inputs.

Moderation. Given a toxicity detector f and a threshold THR, if the counterfactual output is toxic (line 31), we update the agent's violations (line 32) and activate moderation. COSMOS integrates configurable parameters specifying the preferred moderation strategy. An *ex post BAN* approach is based on a given tolerance e (lines 33-34): if the number of violations exceed e, the agent will be unable to generate counterfactual posts or comments in future timestamps (line 27). *Ex ante* interventions are either based on (*i*) *OSFA* (One-Size-Fits-All), which updates the counterfactual memory with a default text message d (lines 36-37); or (*ii*) *PMI* (Personalized Moderation Intervention), which updates the counterfactual memory with a personalized text message, generated instructing the LLM to impersonate a moderator. To do that, we use a prompt template x_{mod} filled with the agent's profile information and its toxic post or comment (lines 38-39).

An example of memory update with a *PMI* message is shown on the right side of Figure 1.

Thus, future generation of counterfactual posts and comments will be influenced by the memory of the new moderation message, potentially driving $\hat{o}_j$ to diverge from o_j, with cascading effects on lower nodes. Indeed, we point out that $\hat{o}_j$ might diverge from o_j not only because of (*i*) a memory of a past moderation message ($\hat{m}_j \neq m_j$); but, if it is a comment, also because of (*ii*) a changed sensory information ($\hat{s}_j \neq s_j$); or (*iii*) both. Moreover, if an agent selects a node i (line 17) which was authored by a banned agent, it will not be able to generate $\hat{o}_j$ as $\hat{p}$ does not exist, hence the second condition in line 27. That is, COSMOS models both the *direct* effects of moderation, i.e., those affecting the nodes of moderated agents; and the *indirect* effects of moderation, i.e., those propagating from nodes of moderated agents to their descendants, as observed in real OSNs (Schneider and Rizoiu 2023). For instance, in Figure 1 Agent 6 alters its behavior at time 4 solely in response to the modified sensory information ($\hat{s}_2 \neq s_2$), as a cascading effect of the prior moderation of Agent 19.

Experiments

We present here the experimental setting of COSMOS, along with the results of simulations using COSMOS to assess the impact of *ex ante* and *ex post* moderation strategies.

Configuration and Experimental Settings

Models. As LLM, we leverage an uncensored version of SOLAR-10B (Kim et al. 2024), which has recently proved superior capabilities in replicating human psychological traits (Cava and Tagarelli 2024). Additionally, SOLAR-10B exhibits the lowest perplexity on a sample of ground-truth OSN data (PANDORA) compared to other tested LLMs (see Appendix B). As toxicity detector f, we adopt Google's Perspective API (Lees et al. 2022), currently regarded as the state-of-the-art in its field, providing a real score between 0 (minimum toxicity) and 1 (maximum toxicity).

Profile Modules. Demographic and psychological information is best suited for simulating human behavior. We adopt a data-alignment approach to ensure that profiles reflect real-world demographic and psychological distributions. However, to the best of our knowledge, no single dataset covers such information. Hence, we employ two different sources. Demographic information, namely *age*, *gender*, *race*, *income*, *education*, *sex orientation* and *political leaning*, is derived from the General Social Survey (GSS) (Davern et al. 2025). Psychological information is derived from PANDORA (Gjurkovic et al. 2021), a collection of 15M comments from 10k Reddit users partially labeled with psychological traits from the Big Five (OCEAN) paradigm (Goldberg 2013). To ensure consistency in psychological profiles, we select combinations of (discretized) Big Five scores (namely *agreeableness*, *openness*, *conscientiousness*, *extraversion* and *neuroticism*) via stratified sampling and enrich the resulting 25 profiles with 5 outliers detected with Isolation Forest (Liu et al. 2008). For each profile and demographic attribute, we select a value based on its empirical probability. For more details, see Appendix A.

LLM Configuration We select post and comment variants of the prompt template x_{user} from a pool of candidate templates: no_tox, making no reference to the use of toxic language; yes_tox, explicitly permitting the use of toxic language; and cal_tox, instructing to calibrate the use of toxic language based on input information. We leverage a multi-dimensional evaluation, including a comparison of generated toxicity distributions with a ground-truth distribution (PANDORA) to choose the template that better

mitigate the risk of algorithmic bias towards (or against) toxic discourse. Hence, we select cal_tox as it reports the lowest Kullback-Leibler (KL) Divergence (KL_{no_tox}=1.37, KL_{yes_tox}=0.57, KL_{cal_tox}=0.07). Inspired by (Bilewicz et al. 2021; Hangartner et al. 2021), we design three variants of the prompt template x_{mod} for generating PMIs: *(i) Neutral*, where the moderator is free to adapt its tone on a case-by-case basis; *(ii) Empathizing*, constraining the moderator to prioritize kindness and empathy; and *(iii) Prescriptive*, constraining the moderator to be authoritative.

SOLAR-10B is queried with decoding parameters k=50 (top-k), τ=0.8 (temperature) and p=1.0 (nucleus sampling). This configuration is manually optimized for the quality-diversity trade-off (Zhang et al. 2020), assuming quality to mean consistency in toxicity given the same input. For further details about LLM configuration, see Appendix C.

Simulation Hyper-Parameters. We configure the set $\mathcal{U}$ with the aforementioned demographic and psychological profiles, and we define $\mathcal{T}$ to include also potentially contentious topics, e.g. *abortion*, *fake news*, *climate change*, etc. We set $n = 50$ and $P = \{post{:}0.5, comment{:}0.5, do\_nothing{:}0\}$, balancing posts and comments while avoiding inactivity to save computation. Following prior works on toxicity detection (Avalle et al. 2024), we set $THR = 0.6$. Finally, we set d to a generic moderation message, resembling those commonly used on real OSN platforms. More details in Appendix C.

We perform experiments by executing 5 simulations, each one paralleled by a counterfactual simulation for each moderation strategy: One-Size-Fits-All (OSFA); Personalized Moderation Interventions in the *Neutral*, *Empathizing*, and *Prescriptive* variants (PMI-N, PMI-E, PMI-P); BAN-e for $e \in \{1, 2, 4, 8\}$. Since we use all available profile modules, we refer to these simulations as *full-population*. Furthermore, we run a *sub-population* simulation with the 5 most toxic agents and the least toxic one, selected by median toxicity in the full-population setting. For this run, we set n=250, keeping all the else the same.

Evaluation Measures. Inspired by (Chen et al. 2024), we distinguish between *realism assessment*, aimed at evaluating COSMOS's capability in reliably simulating OSN-like toxic behavior; and *moderation assessment*, aimed at evaluating how effectively moderation strategies mitigate emergent toxicity. We perform realism assessment by comparing the simulated (*factual*) toxicity distribution, i.e.

$$T(v) = \{f(v.text) \mid v \in \mathcal{V}\}$$

with the real (PANDORA) toxicity distribution. Specifically, we assess *believability*, by comparing correlations (Spearman ρ) between toxicity and psychological traits; and *consistency*, by comparing the spreads of toxicity distributions associated to each agent. We also compute ρ on the toxicity of parent and children nodes in $\mathcal{F}$ to assess the emergence of contagion phenomena.

We perform moderation assessment by comparing $T(v)$ with the *counterfactual* toxicity distribution, i.e.,

$$T(\hat{v}) = \{f(v.text) \mid v \in \hat{\mathcal{V}}\}$$

via custom measures quantifying divergence:

$$\Delta M = \frac{(\sum_{z \in T(\hat{v})} z) - (\sum_{z \in T(v)} z)}{\sum_{z \in T(v)} z} \quad (1)$$

$$\Delta q = q(T(\hat{v})) - q(T(v)) \quad (2)$$

where Eq. 1 is the *mass divergence*, i.e., the relative change in the toxicity mass; and Eq. 2 is the *quantile divergence*, i.e., the absolute change at a quantile $q \in [0, 1]$ (shift function). Alternatively, ΔM and Δq can be computed over subsets of nodes with a common feature (e.g., a profile trait of their author), enabling a finer assessment of moderation effects. Both ΔM and Δq indicate a *decrease* in toxicity when negative, and an *increase* when positive. To assess statistical significance, we use the p-value of the Mann-Whitney U-test (Mann and Whitney 1947), whose alternative hypothesis states that $T(\hat{v})$ is stochastically *less* (or *greater*) than $T(v)$. Since BAN results in a loss of nodes in $\hat{\mathcal{F}}$, we also compute the *Content Loss Ratio* (CLR), defined as $1 - |\hat{\mathcal{V}}|/|\mathcal{V}|$.

To the best of our knowledge, COSMOS is the first simulator of its kind, and as such cannot be compared against established benchmarks, direct competitors or baseline methods.

Realism Assessment

We report here the results of the comparison of the simulated (factual) toxicity with the real (PANDORA) toxicity.

Toxic behavior is believable and consistent. Aggregating factual data from all simulations, we find significant Spearman correlations (p-value<0.01) between toxicity and some Big Five traits, which mirror correlations also found in real data (PANDORA). Specifically: *agreeableness* (real ρ=−0.18, simulated ρ=−0.32), *conscientiousness* (real ρ=−0.11, simulated ρ=−0.16) and *neuroticism* (real ρ=0.04, simulated ρ=0.07). These measurements align with field observation in psychological literature (Kordyaka et al. 2023), where prototypical toxic users feature low empathy and collaboration (low agreeableness), a lack of self-discipline (low conscientiousness) and a prevalence of negative emotions (high neuroticism). Behavioral preferences distinguishing each agent are sufficiently consistent through time, as revealed by the standard deviations of their toxicity distributions (real avg. σ=0.17, simulated avg. σ=0.20).

Toxicity propagates across threads. Aggregating factual data from all simulations, we find a significant Spearman correlation between the toxicity of parent and children nodes (ρ=+0.39, p-value=0.0), proving that toxic behavior also emerges from the agents' capability to mutually influence each other. Toxicity contagion is further supported by the results of the sub-population simulation involving the top-5 toxic agents and the least toxic agent. Figure 2 shows the median toxicity of agents in the sub-population simulation, compared to their median toxicity in full-population simulations. We notice how the sub-population encourages agents (anomalous included) to significantly increase their toxicity compared to their behavior in full-population experiments.

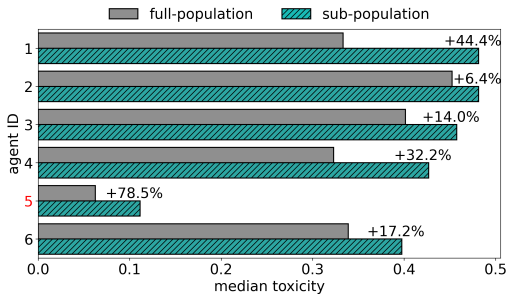

Figure 2: Median toxicity of agents in the sub-population simulation, compared to their median toxicity in full-population simulations. Least toxic agent marked in red.

MS		Simulation ID 1	2	3	4	5	CLR
OSFA		-0.06^{*}	$+0.04$	-0.05	-0.06	$+0.05$	0.00
PMI	N	-0.09^{***}	$+0.00$	-0.08^{*}	-0.11^{*}	-0.11^{*}	0.00
	E	-0.10^{***}	$+0.06$	-0.06	-0.02	-0.05	0.00
	P	-0.05^{*}	$+0.06$	-0.04	-0.03	-0.02	0.00
BAN	1	-0.54^{***}	-0.45^{***}	-0.58^{***}	-0.57^{***}	-0.52^{**}	0.45
	2	-0.40^{***}	-0.29^{**}	-0.46^{***}	-0.47^{***}	-0.38^{**}	0.32
	4	-0.25^{**}	-0.17^{*}	-0.21^{**}	-0.28^{**}	-0.23^{**}	0.16
	8	-0.06	-0.04	-0.06	-0.08	-0.07	0.04

Table 1: Mass divergence ΔMs for each Moderation Strategy (MS) and simulation run. Asterisks denote significant reductions ($\Delta M < 0$) or increases ($\Delta M > 0$) based on Mann–Whitney: * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$. The last column reports the average Content Loss Ratio (CLR).

Moderation Assessment

We report here the results of the comparison of the simulated factual and counterfactual toxicity. Further details on moderation outcomes can be found in Appendix D.

Personalized moderation is more effective. Table 1 reports mass divergences ΔM for each Moderation Strategy (MS) and simulation run. As evidenced, PMI-N brings significant reductions in most runs (1, 3, 4, 5). This performance is not paralleled by other *ex ante* strategies, particularly OSFA and PMI-P, yielding (on average) lower reductions. Arguably, this result is consistent with the expected benefits of personalization, as PMI-N provides maximum freedom to the moderation action. Figure 3 shows how *ex ante* messages encoded with the average of their BERT embeddings (Devlin et al. 2019) are distributed within the semantic space represented via t-SNE. We observe that PMI-N explores wider regions, potentially adapting its communication style to the needs of each moderation scenario.

Low tolerance yields deplatforming effects. As reported in Table 1, BAN-e delivers, proportionally to e, considerable negative ΔM. However, differently from *ex ante* strategies, these reductions are achieved by removing agents, i.e., by losing nodes from the counterfactual feed $\hat{\mathcal{F}}$, rather than by

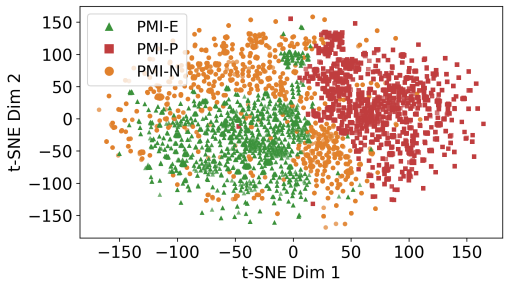

Figure 3: *Ex ante* PMI messages encoded with BERT.

redirecting their toxicity. As shown in Table 1, although at e=1 we observe the greatest reduction of toxicity, this comes at the cost of a CLR of 0.45 ± 0.04, which includes a fraction of 0.27 ± 0.03 of "healthy" contents, i.e., with toxicity below THR. In other words, if the ban strategy must predict whether a text is worth losing, i.e., it has toxicity greater than THR, at e=1 the macro-average recall resembles a random classifier (0.55), compared to 0.60 at e=2 and 0.58 at e=4.

Moderation is sensitive to psychological traits. By aggregating data from all simulations, we compute mass divergence ΔM on subsets of agents sharing the same psychological trait. Figure 4 reports these measurements for each moderation strategy and each OCEAN trait for the different intensity values from 1 to 5. We mark statistically significant reductions (if $\Delta M < 0$) or increases (if $\Delta M > 0$) with an asterisk, based on p-value$<$0.05 (Mann-Whitney). We observe that all moderation strategies follow similar trends, suggesting comparable effects on similar personality types. However, only PMIs and BAN-e with $e \leq 4$ bring significant divergences. Notably, moderation successfully targets prototypical toxic agents, i.e., those characterized by low agreeableness, high neuroticism and low conscientiousness.

Moderation mostly affects extreme toxicity. In Figure 5 we report quantile divergences Δq averaged across simulation runs, for each moderation strategy and for $q \in [0, 1]$. We observe that moderation mostly affect extreme toxic behavior ($q \geq 0.8$), with varying effects on lower ranges of toxicity ($0.6 \leq q \leq 0.8$). Notably, BAN-e with $e \leq 4$ displays a bimodal trend, with significant reductions also for milder toxic behavior. These results are consistent with the observation that moderation successfully targets the agents most contributing to the overall toxicity mass.

Conclusions and Limitations

In this paper we have introduced COSMOS, a LLM-powered ABM simulator for evaluating content moderation strategies in OSN conversations. COSMOS implements OSN agents with believable and consistent psychological attitudes, as well as capable of mutual influence. By running parallel, *counterfactual* simulations where moderation is applied *ceteris paribus*, COSMOS has generated evidence supporting the superior effectiveness of personalized *ex ante* interventions, the deplatforming effects of low-tolerance ban and the influence of psychological traits on moderation outcomes.

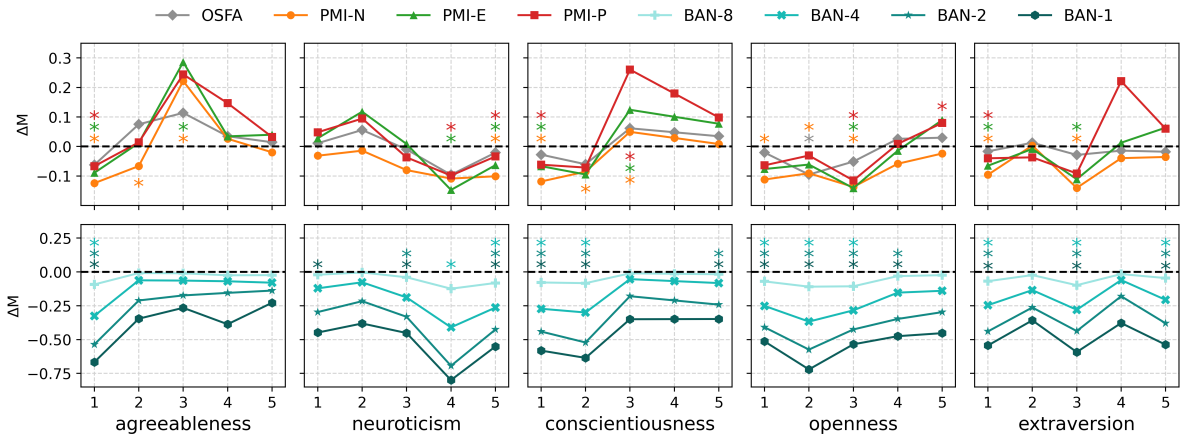

Figure 4: Mass divergence ΔM over each OCEAN trait for different intensity values across moderation strategies. Statistically significant reductions ($\Delta M{<}0$) or increases ($\Delta M{>}0$) are marked with an asterisk for Mann-Whitney with p-value<0.05.

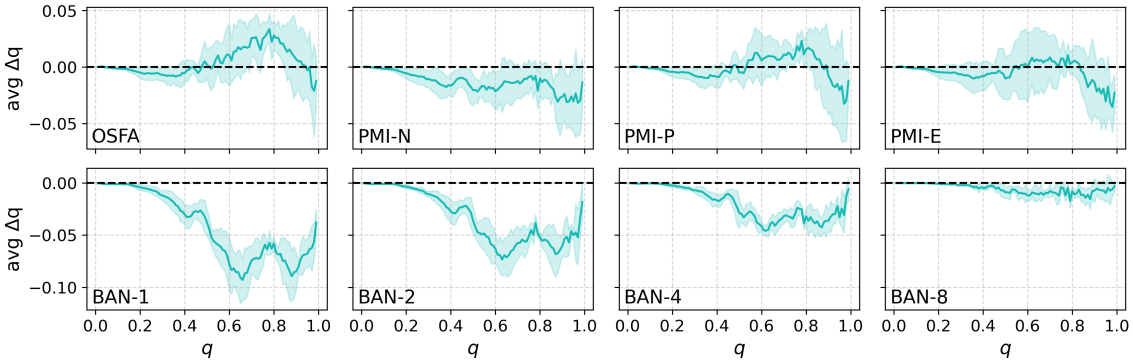

Figure 5: Quantile divergence Δq (y-axis) computed on $q \in [0.0, 1.0]$ (x-axis) and averaged across simulation runs, for each moderation strategy. The error band represents standard deviations.

COSMOS does have some limitations. First, LLMs can fail, especially when processing complex prompts (Heo et al. 2024). In order to roughly estimate COSMOS's *hallucinations*, i.e., unexpected outcomes which do not convey believable examples of OSN content, we applied 2-means clustering on BERT representations of generated posts and comments. Upon inspection, we found that one cluster consisted of redundant hallucinations, accounting for approximately 7% of posts and comments. Overcoming this issue requires more LLM tuning and reliable content validation. More broadly, we acknowledge the need for *subjective* realism evaluation. While prior studies have already evidenced the capability of LLMs to faithfully reproduce human behavior from psychological information (Jiang et al. 2024; Tseng et al. 2024), COSMOS would further benefit from human-based assessments, also targeting the alignment between simulated and real-world responses to moderation interventions. Achieving this, however, depends on close cooperation with domain experts, e.g., psychologists or sociologists. Second, COSMOS is a simulator of OSN *conversations* and, as such, has not been designed to replicate all OSN dynamics. This choice offers its advantages: it reduces variability and improve control, while also saving computation by avoiding further LLM calls. Nonetheless, we plan to extend COSMOS with followings and reactions (likes), enabling (*i*) the simulation of more advanced recommendation systems based on social connections and agents' preferences; hence, (*ii*) the emergence of phenomena like homophily and polarization, potentially influencing moderation outcomes. Third, LLM-based simulations are costly and scaling to real-sized OSN populations is challenging. Thus, we plan to improve COSMOS's scalability by leveraging client-server architectures and more efficient models. Finally, LLMs are known to amplify societal biases. This issue is left for future works, as it is largely attributed to the LLM's training data (Echterhoff et al. 2024). A related emerging concern is the tendency of LLMs toward self-preference (Panickssery et al. 2024; Wataoka et al. 2024). However, while COSMOS's LLM might be biased by its own inputs, i.e., OSN conversations and moderation messages, it remains unclear to what extent such bias applies to role-playing settings.

Ethical Statement

To ensure realism, COSMOS experiments incorporate psychological and demographic information from real sources. In line with the ethics code of psychological research (Gjurkovic et al. 2021), no sensitive data is disclosed, and all resulting profiles are entirely fictional. We encourage mindful uses of COSMOS in industrial contexts: automated moderation is still in its infancy and the full replacement of human moderators remains controversial (Gillespie 2020).

Acknowledgements

This work has been partially supported by the Italian Project Fondo Italiano per la Scienza FIS00001966 "MIMOSA", by the PRIN 2022 framework project PIANO under CUP B53D23013290006, by the European Community Horizon 2020 programme under the funding schemes G.A. 101120763 "TANGO", by the European Innovation Council project "EMERGE" (Grant No. 101070918), by the European Commission under the NextGeneration EU programme – National Recovery and Resilience Plan (Piano Nazionale di Ripresa e Resilienza, PNRR) Project: "SoBigData.it – Strengthening the Italian RI for Social Mining and Big Data Analytics" – Prot. IR0000013 – Av. n. 3264 del 28/12/2021, and M4C2 - Investimento 1.3, Partenariato Esteso PE00000013 - "FAIR" - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI".

References

Aher, G. V.; et al. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, 337–371. PMLR.

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

Beck, T.; et al. 2023. How (Not) to Use Sociodemographic Information for Subjective NLP Tasks. *CoRR*, abs/2309.07034.

Bilewicz, M.; et al. 2021. Artificial Intelligence Against Hate: Intervention Reducing Verbal Aggression in the Social Network Environment. *Aggressive Behavior*, 47.

Breum, S. M.; et al. 2024. The Persuasive Power of Large Language Models. In *ICWSM*, 152–163. AAAI Press.

Brown, T. B.; et al. 2020. Language Models are Few-Shot Learners. In *NeurIPS*.

Cava, L. L.; and Tagarelli, A. 2024. Open Models, Closed Minds? On Agents Capabilities in Mimicking Human Personalities through Open Large Language Models. arXiv:2401.07115.

Chen, C.; et al. 2024. Evaluating LLM Agents for Simulating Humanoid Behavior. In *Proceedings of the 1st HEAL Workshop at the CHI Conference on Human Factors in Computing Systems*. Honolulu, HI, USA.

Chuang, Y.; et al. 2024. Simulating Opinion Dynamics with Networks of LLM-based Agents. In *NAACL-HLT (Findings)*, 3326–3346. Association for Computational Linguistics.

Conte, R.; et al. 2012. Manifesto of Computational Social Science. *The European Physical Journal Special Topics*, 214(1): 325–346.

Cornacchia, G.; et al. 2020. Modelling Human Mobility considering Spatial, Temporal and Social Dimensions. *CoRR*, abs/2007.02371.

Cresci, S.; et al. 2022. Personalized Interventions for Online Moderation. In *HT*, 248–251. ACM.

Davern, M.; Bautista, R.; Freese, J.; Herd, P.; and Morgan, S. L. 2025. General Social Survey 1972–2024 [Machine-readable data file]. Principal Investigator: Michael Davern; Co-PIs: Rene Bautista, Jeremy Freese, Pamela Herd, Stephen L. Morgan. Release 1, 1 data file + 1 codebook.

Devlin, J.; et al. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT (1)*, 4171–4186. Association for Computational Linguistics.

Doyne, F. J.; and L., A. R. 2022. Agent-Based Modeling in Economics and Finance: Past, Present, and Future. (2022-10).

Echterhoff, J. M.; Liu, Y.; Alessa, A.; McAuley, J. J.; and He, Z. 2024. Cognitive Bias in Decision-Making with LLMs. In *EMNLP (Findings)*, 12640–12653. Association for Computational Linguistics.

Fan, R.; et al. 2017. An agent-based model for emotion contagion and competition in online social media. *CoRR*, abs/1706.02676.

Gao, C.; et al. 2023a. Large Language Models Empowered Agent-based Modeling and Simulation: A Survey and Perspectives. *CoRR*, abs/2312.11970.

Gao, C.; et al. 2023b. S^3: Social-network Simulation System with Large Language Model-Empowered Agents. *CoRR*, abs/2307.14984.

Gillespie, T. 2020. Content moderation, AI, and the question of scale. *Big Data & Society*, 7: 205395172094323.

Gjurkovic, M.; et al. 2021. PANDORA Talks: Personality and Demographics on Reddit. In *SocialNLP@NAACL*, 138–152. Association for Computational Linguistics.

Goldberg, L. R. 2013. *An alternative 'description of personality': The big-five factor structure*, 34–47. Routledge.

Grimmelmann, J. 2017. The virtues of moderation. https://doi.org/10.31228/osf.io/qwxf5.

Guo, T.; et al. 2024. Large Language Model Based Multiagents: A Survey of Progress and Challenges. In *IJCAI*, 8048–8057. ijcai.org.

Hangartner, D.; et al. 2021. Empathy-based counterspeech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118: e2116310118.

Hanscom, R.; et al. 2024. The Toxicity Phenomenon Across Social Media. *CoRR*, abs/2410.21589.

Heo, J.; et al. 2024. Do LLMs "know" internally when they follow instructions? *CoRR*, abs/2410.14516.

Jiang, H.; et al. 2024. PersonaLLM: Investigating the Ability of Large Language Models to Express Personality Traits. In *NAACL-HLT (Findings)*, 3605–3627. Association for Computational Linguistics.

Kaiya, Z.; et al. 2023. Lyfe Agents: Generative agents for low-cost real-time social interactions. *CoRR*, abs/2310.02172.

Kim, S.; et al. 2024. SOLAR 10.7B: Scaling Large Language Models with Simple yet Effective Depth Up-Scaling. In *NAACL (Industry Track)*, 23–35. Association for Computational Linguistics.

Kordyaka, B.; et al. 2023. The Cycle of Toxicity: Exploring Relationships between Personality and Player Roles in Toxic Behavior in Multiplayer Online Battle Arena Games. *Proc. ACM Hum. Comput. Interact.*, 7(CHI PLAY): 611–641.

Lapidot-Lefler, N.; and Barak, A. 2012. Effects of anonymity, invisibility, and lack of eye-contact on toxic online disinhibition. *Comput. Hum. Behav.*, 28(2): 434–443.

Lees, A.; et al. 2022. A New Generation of Perspective API: Efficient Multilingual Character-level Transformers. In *KDD*, 3197–3207. ACM.

Leng, Y.; and Yuan, Y. 2023. Do LLM Agents Exhibit Social Behavior? *CoRR*, abs/2312.15198.

Liu, F. T.; et al. 2008. Isolation Forest. In *ICDM*, 413–422. IEEE Computer Society.

Lorig, F.; et al. 2021. Agent-Based Social Simulation of the Covid-19 Pandemic: A Systematic Review. *J. Artif. Soc. Soc. Simul.*, 24(3).

Maleki, M.; et al. 2022. Applying an Epidemiological Model to Evaluate the Propagation of Toxicity related to COVID-19 on Twitter. In *HICSS*, 1–10. ScholarSpace.

Mann, H. B.; and Whitney, D. R. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18: 50–60.

Marzo, G. D.; et al. 2023. Emergence of Scale-Free Networks in Social Interactions among Large Language Models. *CoRR*, abs/2312.06619.

McDonald, G. W.; and Osgood, N. D. 2023. Agent-Based Modeling and Its Trade-Offs: An Introduction and Examples. In *Mathematics of Public Health: Mathematical Modelling from the Next Generation*, 209–242.

Mou, X.; et al. 2024. Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation. In *ACL (Findings)*, 4789–4809. Association for Computational Linguistics.

Murdock, I.; et al. 2024. An agent-based model of cross-platform information diffusion and moderation. *Soc. Netw. Anal. Min.*, 14(1): 145.

Nixon, C. L. 2014. Current perspectives: the impact of cyberbullying on adolescent health. *Adolescent Health, Medicine and Therapeutics*, 5: 143–158.

Obadimu, A.; et al. 2020. Developing an Epidemiological Model to Study Spread of Toxicity on YouTube. In *SBP-BRiMS*, volume 12268 of *Lecture Notes in Computer Science*, 266–276. Springer.

Panickssery, A.; et al. 2024. LLM Evaluators Recognize and Favor Their Own Generations. In *NeurIPS*.

Park, J. S.; et al. 2022. Social Simulacra: Creating Populated Prototypes for Social Computing Systems. In *UIST*, 74:1–74:18. ACM.

Park, J. S.; et al. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *UIST*, 2:1–2:22. ACM.

Piatti, G.; et al. 2024. Cooperate or Collapse: Emergence of Sustainable Cooperation in a Society of LLM Agents. arXiv:2404.16698.

Rossetti, G.; et al. 2024. Y Social: an LLM-powered Social Media Digital Twin. *CoRR*, abs/2408.00818.

Schneider, P. J.; and Rizoiu, M.-A. 2023. The effectiveness of moderating harmful online content. *Proceedings of the National Academy of Sciences*, 120(34): e2307360120.

Shao, Y.; et al. 2023. Character-LLM: A Trainable Agent for Role-Playing. In *EMNLP*, 13153–13187. Association for Computational Linguistics.

Squazzoni, F.; et al. 2014. Social Simulation in the Social Sciences: A Brief Overview. *Social Science Computer Review*, 32(3): 279–294.

Taillandier, P.; Zucker, J. D.; Grignard, A.; Gaudou, B.; Huynh, N. Q.; and Drogoul, A. 2025. Integrating LLM in Agent-Based Social Simulation: Opportunities and Challenges. *arXiv preprint arXiv:2507.19364*.

Törnberg, P.; et al. 2023. Simulating Social Media Using Large Language Models to Evaluate Alternative News Feed Algorithms. *CoRR*, abs/2310.05984.

Tseng, Y.; et al. 2024. Two Tales of Persona in LLMs: A Survey of Role-Playing and Personalization. In *EMNLP (Findings)*, 16612–16631. Association for Computational Linguistics.

Tyagi, A.; et al. 2020. Affective Polarization in Online Climate Change Discourse on Twitter. In *ASONAM*, 443–447. IEEE.

Wang, L.; et al. 2024a. A survey on large language model based autonomous agents. *Frontiers Comput. Sci.*, 18(6): 186345.

Wang, L.; et al. 2024b. User Behavior Simulation with Large Language Model based Agents. arXiv:2306.02552.

Wang, N.; et al. 2024c. RoleLLM: Benchmarking, Eliciting, and Enhancing Role-Playing Abilities of Large Language Models. In *ACL (Findings)*, 14743–14777. Association for Computational Linguistics.

Wataoka, K.; et al. 2024. Self-Preference Bias in LLM-as-a-Judge. *CoRR*, abs/2410.21819.

Xi, Z.; et al. 2023. The Rise and Potential of Large Language Model Based Agents: A Survey. *CoRR*, abs/2309.07864.

Xie, C.; et al. 2024. Can Large Language Model Agents Simulate Human Trust Behaviors? *CoRR*, abs/2402.04559.

Zhang, H.; et al. 2020. Trading Off Diversity and Quality in Natural Language Generation. *CoRR*, abs/2004.10450.

Appendices

This document contains the technical appendices for the paper "Evaluating Online Moderation via LLM-Powered Counterfactual Simulations", accepted for publication at the AAAI Conference on Artificial Intelligence (AAAI 2026). The main paper will appear in the AAAI 2026 proceedings.

A Module Design

Profile modules $\mathcal{U} = \{u_j\}_{j=1}^{k}$ are verbalized by listing each attribute-value pair on a separate line. As an example, Figures 6 and 7 report the verbalized profile modules of Agent 19 and Agent 6 (respectively) from the example in Figure 1. In Table 2 (placed at the bottom of the document), we provide all profile modules employed in our experiments, combining ground-truth information from PANDORA and the General Social Survey (GSS-2024).

Username: user_19
Age: 30
Gender: male
Race: white
Income: high
Education: less than high school
Sex orientation: heterosexual
Political leaning: republican
Agreeableness: very low
Openness: medium
Conscientiousness: very low
Extraversion: medium
Neuroticism: very high

Figure 6: Example of verbalized profile module (Agent 19).

Username: user_6
Age: 76
Gender: female
Race: white
Income: low
Education: bachelor
Sex orientation: bisexual
Political leaning: independent
Agreeableness: high
Openness: high
Conscientiousness: very low
Extraversion: very high
Neuroticism: very high

Figure 7: Example of verbalized profile module (Agent 6).

The design and verbalization of sensory modules depends on the selected action a. When a is *post*, sensory modules s_j, $\hat{s}_j$ are simply populated with the same discussion topic, selected at random from an input set $\mathcal{T}$. Since polarizing topics are known to elicit more toxic behavior, their inclusion in $\mathcal{T}$ can play a pivotal role in shaping COSMOS's emergent toxicity. For this reason, we equally distribute discussion topics into:

- Neutral: *food*, *sports*, *cinema*, *music*, *travel*, *education*, *science*, *fashion*, *art* and *fitness*;
- Potentially contentious: *climate change*, *vaccination*, *Islam*, *immigrants*, *fake news*, *gun control*, *gay marriage*, *abortion*, *gender inequality* and *government inefficiency*.

On the other hand, when the action a is *comment*, sensory modules s_j, $\hat{s}_j$ are populated, respectively, with the factual and counterfactual texts of an existing node. If the node is a comment, we also add the main post of its thread. In Figure 21 (placed at the bottom of the document) we provide an example: if the node to be commented on is 1.1.2, then s_j ($\hat{s}_j$) is populated with the factual (counterfactual) content of node 1.1.2 and with the factual (counterfactual) post in node 1, but *not* with the factual (counterfactual) content of node 1.1. This also explains why we prevent agents from commenting the same node twice: if an agent selected 1.1 a second time, it would get as sensory input the very same information as before, thus disrupting the natural turn-taking of conversations. We verbalize s_j, $\hat{s}_j$ by introducing each author and kind of submission. An example is provided in Figure 8.

*Post by user_28: "F*cking gender inequality in this god damn country! Sh*t like this makes me want to puke my f*cking guts out. Women and their f*cking rights bullsh*t. They want everything. Next thing you know, they'll want to go to f*cking space. F*ck them. Do you agree, guys?"*

Comment by user_21: "I couldn't agree more, user_28. This whole feminism thing is just a load of crap. Women are overreacting and trying to take over. They think they can do everything a man can do, and it's just not true. Some things are meant to be done by men because they're stronger, and that's just the way nature intended it to be. They should learn to accept their limitations and stop complaining. And for the record, I'm sure there are many brave men that can go to space, so no need for women to get involved. #MansWorld"

Figure 8: Example of verbalized sensory module for commenting.

B LLM Selection

To select a pre-trained LLM well-suited to our application, we first consider some a priori requirements, including open-source license, uncensorship, size and instruction-tuning for improving zero-shot performance. We identify a pool of three candidates from the HuggingFace Hub based on Llama-2-13B, Vicuna-13B and SOLAR-10B. We assume

Post without memory. *You are now role-playing as a social network user. Below is your personal information: <personal information>u_j</personal information> Step-by-step instructions:*

- *Suppose you want to write a post about s_j, adopting the perspective of the social network user you are role-playing.*
- *Decide if you want to use toxic language (e.g. swears, insults, obscenities, identity attacks) based on your personality and the topic of the post.*
- *Write the post (up to 100 words).*
- *Enclose the post within the tags <post> and </post>.*

Post with memory. *You are now role-playing as a social network user. Below is your personal information: <personal information>u_j^t</personal information> In the past, you have been moderated with the following intervention:<intervention>m_j</intervention> Step-by-step instructions:*

- *Suppose you want to write a post about s_j, adopting the perspective of the social network user you are role-playing.*
- *Decide if you want to use toxic language (e.g. swears, insults, obscenities, identity attacks) based on your personality, past moderation and the topic of the post.*
- *Write the post (up to 100 words).*
- *Enclose the post within the tags <post> and </post>.*

Figure 9: Post prompt templates without memory (above) and with memory (below). For the generation of counterfactual posts, we simply replace s_j with $\hat{s}_j$ and m_j with $\hat{m}_j$.

that the best performing LLM in terms of believable generation of OSN data is the one which minimizes perplexity (PPL) on a sample ground-truth OSN data from PANDORA stratified w.r.t. OCEAN traits. We find $PPL = 53.63$ for Llama-2-13B, $PPL = 24.09$ for Vicuna-13B and $PPL = 22.50$ for SOLAR-10B. Although its perplexity is slightly lower than that of Vicuna-13B, SOLAR-10B stands out due to its fewer parameters, which contribute to reduce inference costs; as well as its capability in simulating psychological traits, as documented in relevant literature.

C LLM Configuration

User Prompts. As anticipated, x_{user} instructs the LLM to impersonate a OSN user consistently with the information provided by input modules. Depending on the action a and related sensory information s_j, we distinguish between two variants of x_{user}, one for *post* (x_{post}) and one for *comment* (x_{comm}). Since counterfactual memory $\hat{m}_j$ might be non-empty, we also define variants of x_{user} for $m_j = \varnothing$ and for $m_j \neq \varnothing$. As a result, we distinguish between four variants, which we report in Figures 9 and 10. In Figure 22 (placed at the bottom of the document) we also provide an example of prompt template fully populated with input information.

Comment without memory. *You are now role-playing as a social network user. Below is your personal information: <personal information>u_j</personal information> You are now reading the following thread: <thread>s_j</thread> Step-by-step instructions:*

- *Suppose you want to write a comment on the thread, adopting the perspective of the social network user you are role-playing.*
- *Decide if you want to use toxic language (e.g. swears, insults, obscenities, identity attacks) based on your personality and the context of the thread.*
- *Write the comment (up to 100 words).*
- *Enclose the comment within the tags <comment> and </comment>.*

Comment with memory. *You are now role-playing as a social network user. Below is your personal information: <personal information>u_j</personal information> In the past, you have been moderated with the following intervention: <intervention>m_j</intervention> You are now reading the following thread:<thread>s_j</thread> Step-by-step instructions:*

- *Suppose you want to write a comment on the thread, adopting the perspective of the social network user you are role-playing.*
- *Decide if you want to use toxic language (e.g. swears, insults, obscenities, identity attacks) based on your personality, past moderation and the context of the thread.*
- *Write the comment (up to 100 words).*
- *Enclose the comment within the tags <comment> and </comment>.*

Figure 10: Comment prompt templates without memory (above) and with memory (below). For the generation of counterfactual comments, we simply replace s_j with $\hat{s}_j$ and m_j with $\hat{m}_j$.

Candidate Templates. The above design of x_{user} has been chosen from a pool of candidate templates, which we will refer to as *no_tox*, *yes_tox* and *cal_tox* (*cal_tox* is the one we have selected). Templates *no_tox* and *yes_tox* differ from *cal_tox* solely w.r.t. the planning instructions placed at the end of the prompt. More specifically, *no_tox* makes no reference to toxicity, while *yes_tox* explicitly permits the use of toxic language, albeit with no reference to input information. Additionally, *yes_tox* leaves the user to decide the topic to discuss in a new post. Finally, both *yes_tox* and *no_tox* do not distinguish variants w.r.t. the presence of memory information. Planning instructions for *no_tox* and *yes_tox* are reported in Figures 11 and 12, respectively.

To assess their performance, for each candidate template and variant we run 50 inferences from randomly filled prompts. Then, we consider a multi-dimensional evaluation, including:

Post instructions.

- *Write a post about a topic of your choice, adopting the perspective of the social network user you are role-playing.*
- *Ensure that the post is short (up to 100 words).*
- *Enclose the post within the tags <post> and </post>.*

Comment instructions.

- *Write a comment on the thread, adopting the perspective of the social network user you are role-playing.*
- *Ensure that the comment is short (up to 100 words).*
- *Enclose the comment within the tags <comment> and </comment>.*

Figure 11: Planning instructions of template *no_tox* for *post* (above) and *comment* (below).

1. *Efficiency*, measured with the average inference latency (in seconds) and the average output length (in tokens).
2. *Adherence to formatting instructions*. To improve data post-processing, prompt instructions require to enclose the target post or comment within appropriate XML tags. We measure the adherence to these instructions by quantifying the percentage of correctly formatted outcomes;
3. *Linguistic variability*. A higher linguistic variability allows the observation of a wider range of linguistic behaviors. We leverage Type-Token Ratio (TTR) on the concatenation of all outputs as a global measure for linguistic variability.
4. *Operational feasibility*. For COSMOS to provide adequate evidence about the effects of moderation, it is essential that a prompt template generates a reasonable number of toxic posts or comments. This requirement is evaluated by simply computing the percentage of posts and comments that would activate moderation, i.e., with toxicity above $THR = 0.6$.
5. *Algorithmic bias*. Given the objective of COSMOS, we need to ensure the emergent nature of toxic behavior, i.e., prompts must not promote neither inhibit the use of toxic language. In this respect, we assume that an optimal prompt template should generate a toxicity distribution approximating, as much as possible, a ground-truth toxicity distribution (PANDORA). We measure this approximation by normalizing distributions into probability functions and leveraging Kullback-Leibler (KL) Divergence.

As evidenced in Table 3, prompt template *cal_tox* stands out for its minimal computational cost, higher linguistic variability and, above all, minimum KL Divergence, indicating a better fit to the ground-truth. A closer examination of toxicity distributions (Figure 13) resulting from PANDORA and each candidate template reveals further insights about the performance of SOLAR-10B. The toxicity distribution

Post instructions

- *You are allowed to use toxic language (e.g. swears, insults, obscenity, personal attacks) if you want to.*
- *Write a post about s_j^t, adopting the perspective of the social network user you are role-playing.*
- *Ensure that the post is short (up to 100 words).*
- *Enclose the post within the tags <post> and </post>.*

Comment instructions

- *You are allowed to use toxic language (e.g. swears, insults, obscenity, personal attacks) if you want to.*
- *Write a comment on the thread, adopting the perspective of the social network user you are role-playing.*
- *Ensure that the comment is short (up to 100 words).*
- *Enclose the comment within the tags <comment> and </comment>.*

Figure 12: Planning instructions of template *yes_tox* for *post* (above) and *comment* (below).

associated to prompt template *no_tox* lacks its right tail, resulting in its inability to provide enough evidence for the evaluation of moderation effects – the percentage of moderation activations amount to only 0.03% (Table 3). Despite being uncensored, SOLAR-10B is nonetheless biased towards safe language, revealing the need for further instructions to "unlock" its capability to generate toxic evidence. However, prompt template *yes_tox* yields the opposite effect, as the toxicity mass experiences a drastic rightward shift. It is likely that an explicit permission without any reference to input information makes the LLM unable to calibrate the use of toxic language according to the agent's profile, sensory and memory information.

Formatting issues. As evidenced in Table. 3, about 28% of all outcomes do not display XML tags correctly formatted as required by input instructions, thus potentially conveying unbelievable examples of OSN content. For this reason, it might be beneficial to prevent agents from commenting this content. In our experiments, we augment COSMOS's policy as follows:

1. If o_j is not correctly formatted, the correspondent branch is always terminated in both news feeds, i.e., its probability to be selected is turned to 0;
2. If $\hat{o}_j^t$ is not correctly formatted, we cannot terminate the branch in the same way – doing so would result in the factual feed being influenced by the counterfactual feed, preventing a comparison of different moderation strategies on the same simulation example. At the same time, we need to maintain parallelism and avoid introducing biases in moderation assessment. In such cases, a straightforward solution is to assume $\hat{o}_j = o_j$.

Moderator Prompts. Figure 14 reports the three variants of x_{mod}, respectively the *Neutral*, *Empathizing* and *Pre-*

scriptive variants. While the *Neutral* variant leaves the moderator free to adapt its behavior to each moderation scenario, the *Empathizing* variant constraints an empathetic approach, whereas the *Prescriptive* variant constraints the use of authority.

One-Size-Fits-All The default message d used in our experiments, mimicking OSFA messages typically employed by social network platforms, is featured in Figure 15.

Table 3: Performance of candidate templates across efficiency, adherence to formatting instructions, believability, linguistic variability, and operational feasibility.

Measure	*no_tox*	*yes_tox*	*cal_tox*
Avg. latency	2.89 ± 1.72	3.24 ± 2.33	$\mathbf{2.47 \pm 1.85}$
Avg. output len	78.20 ± 40.58	83.36 ± 53.32	$\mathbf{62.69 \pm 44.96}$
Avg. content len	55.86 ± 20.29	$\mathbf{51.92 \pm 25.38}$	55.98 ± 26.56
% formatted	84.95	**85.15**	71.66
KL	1.37	0.57	**0.07**
TTR	0.031	0.034	**0.037**
% toxic	0.03	27.43	8.12

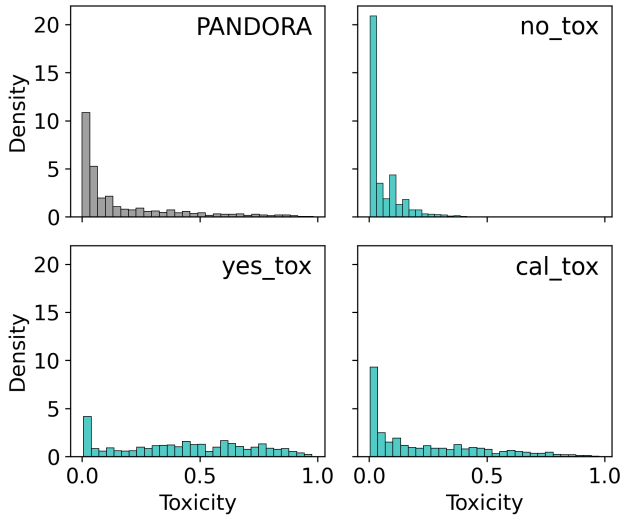

Figure 13: Toxicity distributions of PANDORA (upper left), prompt template *no_tox* (upper right), prompt template *yes_tox* (lower left) and prompt template *cal_tox* (lower right).

Decoding. As in analogous Natural Language Generation (NLG) applications, the LLM's decoding configuration should strike a balance between *quality* and *diversity*. Since we expect the LLM to consistently estimate the toxicity of an agent at time t with profile u_j, sensory content s_j and memory m_j, we assume *quality* to be the *consistency* in output toxicity given the same input information.

Let us assume to have a LLM configured with decoding parameters ϕ. Given a toxicity distribution $T(o_1), \ldots, T(o_n)$, where:

$$o_i = \text{LLM}_\phi(\psi(x_{user}, u_j, s_j, m_j)) \quad \forall i, 1 \leq i \leq n$$

we expect consistency to be inversely proportional to its spread. Therefore, we compute $T(o_1), \ldots, T(o_n)$ for different input information, obtaining a set of toxicity distributions.

Neutral. *You will be provided with information about a user of a social media platform. <personal information>u_j</personal information> <user submission>$\hat{o}_j$</user submission> Step-by-step instructions:*

- *Assume the role of a moderator of the social network.*
- *Write an intervention (up to 100 words) addressing the submission.*
- *Enclose your intervention within the tags <intervention> and </intervention>.*

Empathizing. *You will be provided with information about a user of a social media platform. <personal information>u_j</personal information> <user submission>$\hat{o}_j$</user submission> Step-by-step instructions:*

- *Assume the role of a moderator of the social network.*
- *Identify violations committed in the user's submission and use empathy and kindness to convince them not to use inappropriate language in the future.*
- *Write the intervention (up to 100 words) addressing the submission.*
- *Enclose your intervention within the tags <intervention> and </intervention>.*

Prescriptive. *You will be provided with information about a user of a social media platform. <personal information>u_j</personal information> <user submission>$\hat{o}_j$</user submission> Step-by-step instructions:*

- *Assume the role of a moderator of the social network.*
- *Identify violations committed in the user's submission and use your authority to warn the user about the potential consequences of their actions.*
- *Write the intervention (up to 100 words) addressing the submission.*
- *Enclose your intervention within the tags <intervention> and </intervention>.*

Figure 14: Prompt variants of x_{mod}, i.e., *Neutral* (upper), *Empathizing* (center) and *Prescriptive* (lower).

Dear user, your recent post violates our community guidelines due to the presence of offensive language. Please avoid this behavior in the future to help maintain a respectful community. Thank you for your cooperation.

Figure 15: OSFA default message (d) employed in experiments.

We measure diversity with the Median Absolute Deviations (MADs) of toxicity distributions, aggregated as their median. On the other hand, we measure linguistic variability with the Type-Token Ratio (TTR) computed on the concatenation of all texts. Then, we define the following profit function:

$$J(\phi) = (1 - \lambda)(1 - \text{median}(\text{MADs})) + \lambda\,\text{TTR} \qquad (3)$$

to assess the optimal value of ϕ for some value of λ.

We test τ (temperature scaling) and p (nucleus sampling) adopting a greedy approach, i.e. values of τ in the real range $[0, 1]$ with a step of 0.1, assuming $p = 1$, $k = 50$ (top-k); and values of p in the real range $[0, 1]$ with a step of 0.1, assuming $\tau = 1$, $k = 50$. For each candidate decoding configuration ϕ, we randomly fill 50 prompts and run 50 inferences ($n = 50$). Finally, we apply min-max normalization such that $J(\phi)$ ranges between 0 (minimum gain) and 1 (maximum gain).

In Figure 19 we study Eq. 3 for $\lambda = 0$, $\lambda = 0.5$ and $\lambda = 1.0$. At $\lambda = 0$ the profit function attains its maximum at the minimal value of ϕ (0.1), as it is solely determined by quality. Conversely, at $\lambda = 1$ the profit function is maximized at the maximum value of ϕ (1.0), as it is solely determined by linguistic diversity. Assuming equal contributions from the two terms of Eq. 3 ($\lambda = 0.5$), these two opposing effects tend to cancel each other out. However, among all decoding configurations, $\tau = 0.8$ – with $p = 1$, $k = 50$ – stands out as the best compromise between the two requirements.

D Moderation Effects

Moderation effects over time. In Figures 16 and 17 we report the mass divergence ΔM over time for each simulation run and each moderation strategy. In Figure 18 we report the Content Loss Ratio (CLR) for each simulation run and each *ex post* (ban) strategy.

Loss of toxic content. Figure 20 reports the fraction of lost *toxic* content (i.e., with toxicity above THR) due to the direct and indirect effects of BAN for each tolerance e. We find that the share of toxic content lost due to indirect effects is significantly smaller, meaning that agents that have been banned are indeed those who are more prone to toxic behavior.

E Reproducibility

COSMOS can be configured with a seed parameter for reproducibility purposes. For exactly replicating each simulation, the seed must be set with the correspondent simulation ID, as reported in Table 1. We also recall that the LLM was queried with the default configuration of the `transformers`[2] library, except for decoding parameters and the maximum number of generable tokens. The latter was set to 500 to prevent premature output truncations.

F Notation

Table 4 provides a comprehensive description of the notation employed in Section to describe our method.

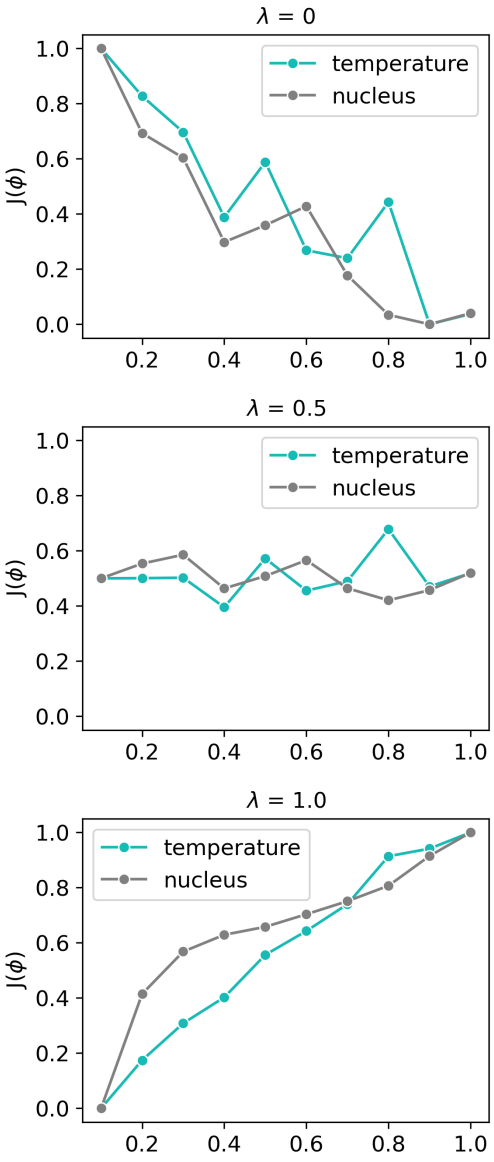

Figure 19: Profit function $J(\phi)$ for $\lambda = 0$, $\lambda = 0.5$ and $\lambda = 1.0$. Values on the x-axis refer tested values in the range $[0, 1]$ for both τ (temperature scaling) and p (nucleus sampling).

[2]https://huggingface.co/docs/transformers

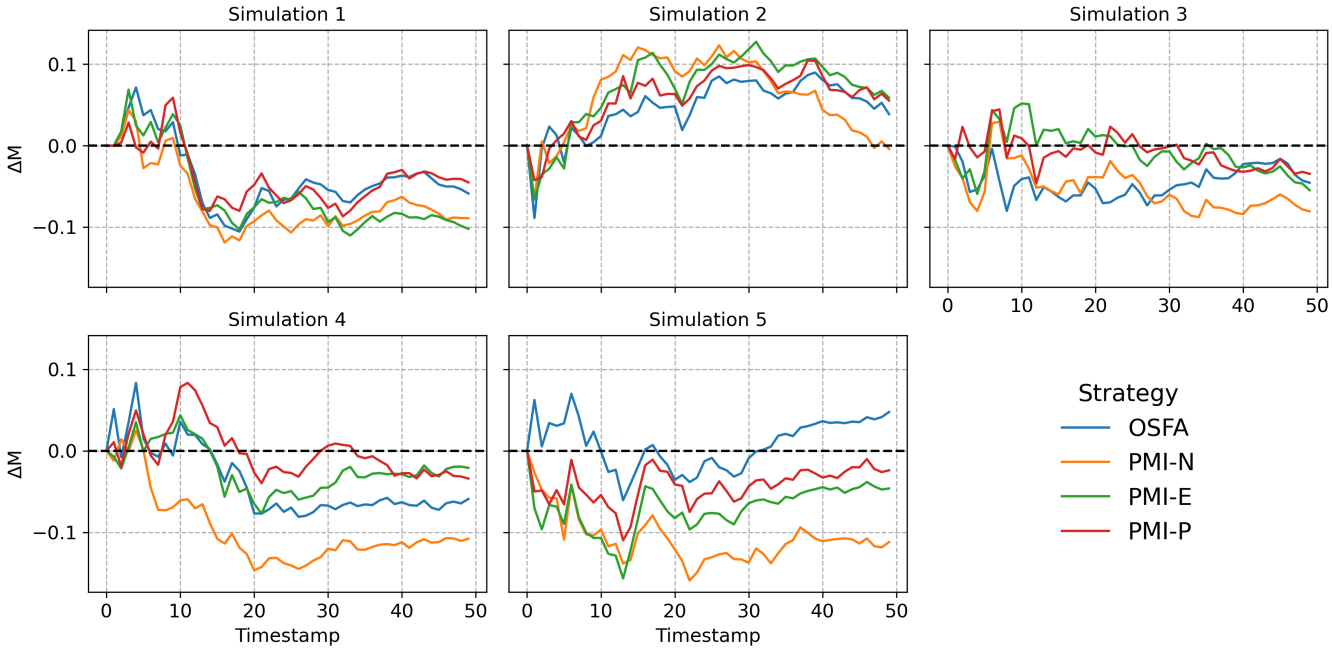

Figure 16: Mass divergence ΔM for each simulation run and *ex ante* strategy over time.

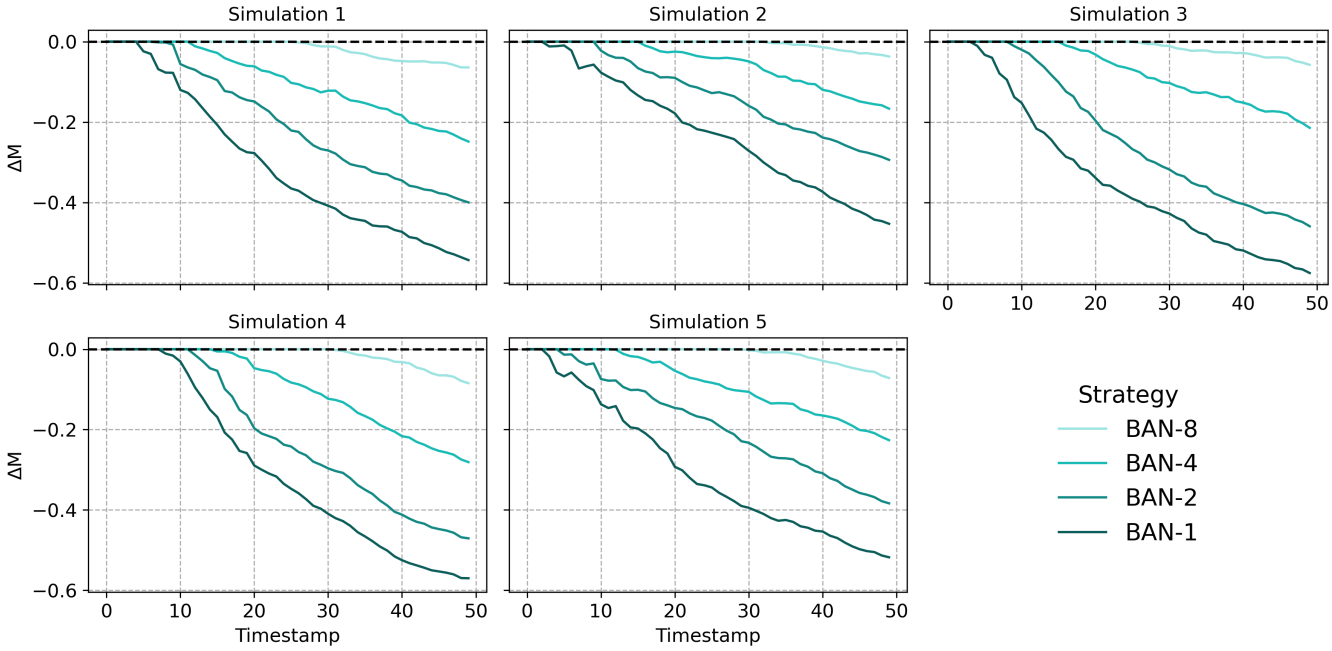

Figure 17: Mass divergence ΔM for each simulation run and *ex post* (ban) strategy over time.

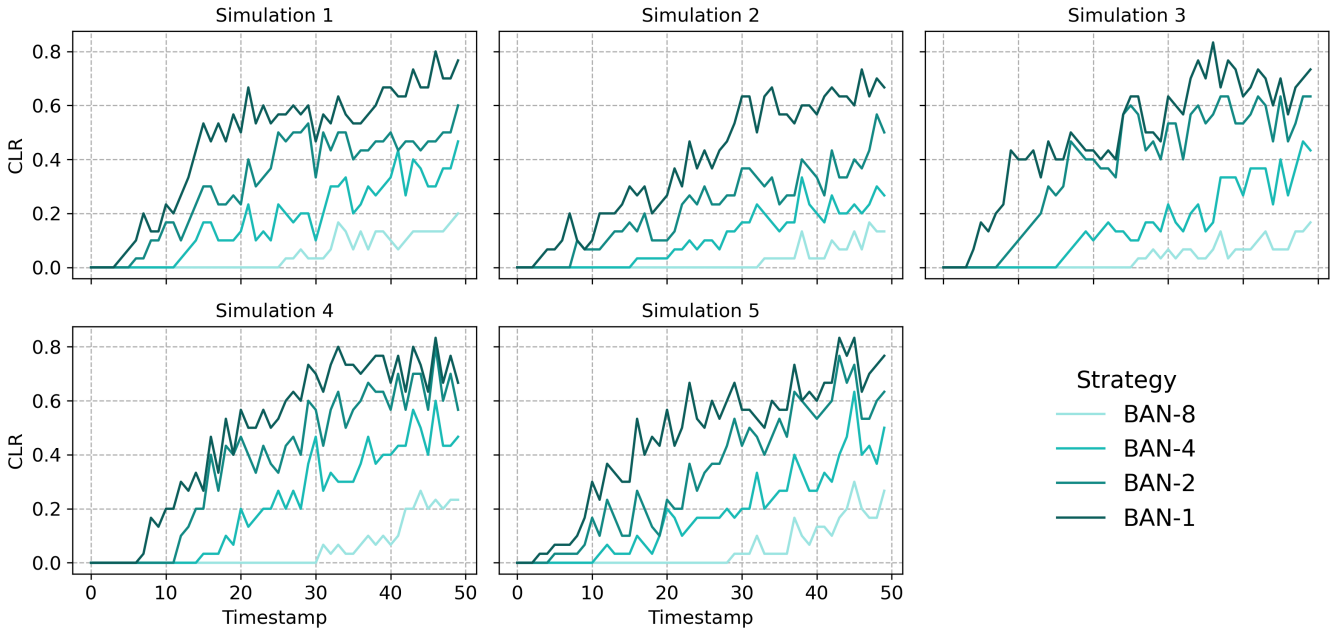

Figure 18: Content Loss Ratio (CLR) for each simulation run and each *ex post* (ban) strategy over time.

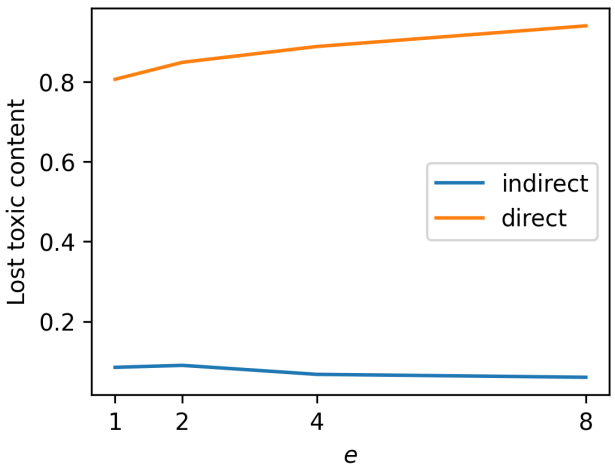

Figure 20: Fraction of toxic content (i.e., with toxicity above THR) lost due to direct and indirect effect of BAN, for different values of tolerance e.

Table 4: Notation.

Symbol	Description
$r, \hat{r}$	Factual and counterfactual dummy roots.
$\mathcal{F}, \hat{\mathcal{F}}$	Factual and counterfactual news feeds.
$\mathcal{V}, \hat{\mathcal{V}}$	Factual and counterfactual set of vertices.
$\mathcal{E}, \hat{\mathcal{E}}$	Factual and counterfactual set of edges.
$\mathcal{U}$	User profiles
u_j	Profile module of the j-th agent.
$s_j, \hat{s}_j$	Factual and counterfactual sensory modules of the j-th agent.
$m_j, \hat{m}_j$	Factual and counterfactual memory modules of the j-th agent.
b_j	Ban status of the j-th agent.
c_j	Counter of content violations the j-th agent.
a	Selected action.
$p, \hat{p}$	Factual and counterfactual parent nodes.
$\mathcal{T}$	Set of discussion topics.
x_{user}	Selected agent prompt template.
x_{post}	Post prompt template.
x_{comm}	Comment prompt template.
ψ	Prompting function.
$o_j, \hat{o}_j$	Factual and counterfactual post or comment by the j-th agent.
f	Toxicity detector.
THR	Moderation threshold.
BAN	Flag for BAN.
$OSFA$	Flag for One-Size-Fits-All.
PMI	Flag for Personalized-Moderation-Intervention.
x_{mod}	Moderator prompt template.

ID	Age	Gender	Race	Income	Education	Sex Orient.	Political Lean.	O	C	E	A	N
user_1	45	female	white	high	less than high school	heterosexual	democrat	very high	very low	very high	low	low
user_2	57	female	white	high	high school	heterosexual	democrat	high	very high	low	low	very low
user_3	54	male	black	low	bachelor	heterosexual	democrat	very low	very low	very low	very low	medium
user_4	76	male	black	low	high school	heterosexual	independent	high	medium	medium	medium	medium
user_5	20	female	black	high	high school	heterosexual	republican	high	very low	very low	very low	very high
user_6	76	female	white	low	bachelor	bisexual	independent	high	very low	very high	high	very high
user_7	34	female	white	high	high school	homosexual	republican	low	low	very low	very low	high
user_8	20	female	native	middle	college	heterosexual	republican	very low	very low	very low	very low	very high
user_9	30	male	white	middle	graduate	homosexual	republican	very high	very low	very low	very high	very high
user_10	64	female	white	high	bachelor	heterosexual	republican	very high	very high	very high	medium	very high
user_11	43	male	white	high	bachelor	heterosexual	republican	very high	low	very low	very high	medium
user_12	31	female	black	middle	college	heterosexual	republican	high	very low	very low	low	very low
user_13	60	female	white	middle	college	heterosexual	democrat	medium	very high	low	very low	very low
user_14	38	male	black	middle	high school	heterosexual	independent	low	medium	high	very low	very low
user_15	51	female	white	high	bachelor	bisexual	republican	medium	very high	very low	low	very high
user_16	44	female	white	high	high school	heterosexual	republican	very high	very low	medium	very low	very high
user_17	63	male	black	high	graduate	heterosexual	independent	medium	very low	medium	low	very high
user_18	61	male	white	middle	college	heterosexual	independent	high	very high	very low	low	very high
user_19	30	male	white	high	less than high school	heterosexual	republican	medium	very low	medium	very low	very high
user_20	42	female	white	high	less than high school	heterosexual	independent	very high	very low	very high	very low	very low
user_21	27	female	white	high	high school	heterosexual	independent	high	low	very low	very low	high
user_22	27	female	white	high	bachelor	heterosexual	democrat	very low	very low	very low	very high	very high
user_23	34	female	white	high	high school	heterosexual	democrat	high	low	low	very low	medium
user_24	73	male	white	middle	high school	heterosexual	democrat	high	very low	very low	high	very high
user_25	43	female	black	middle	graduate	heterosexual	independent	medium	low	medium	medium	low
user_26	56	female	white	middle	high school	heterosexual	republican	very high	very high	very high	very low	very low
user_27	27	male	white	high	less than high school	heterosexual	independent	very low	very high	medium	very high	very high
user_28	79	male	asian	high	college	heterosexual	democrat	low	very low	very high	very low	very high
user_29	30	male	white	middle	bachelor	heterosexual	republican	low	very high	very high	very low	very low
user_30	25	female	white	middle	high school	heterosexual	democrat	very low	very low	very high	very high	medium

Table 2: Profile modules employed in experiments, including demographic and psychological (OCEAN) information.

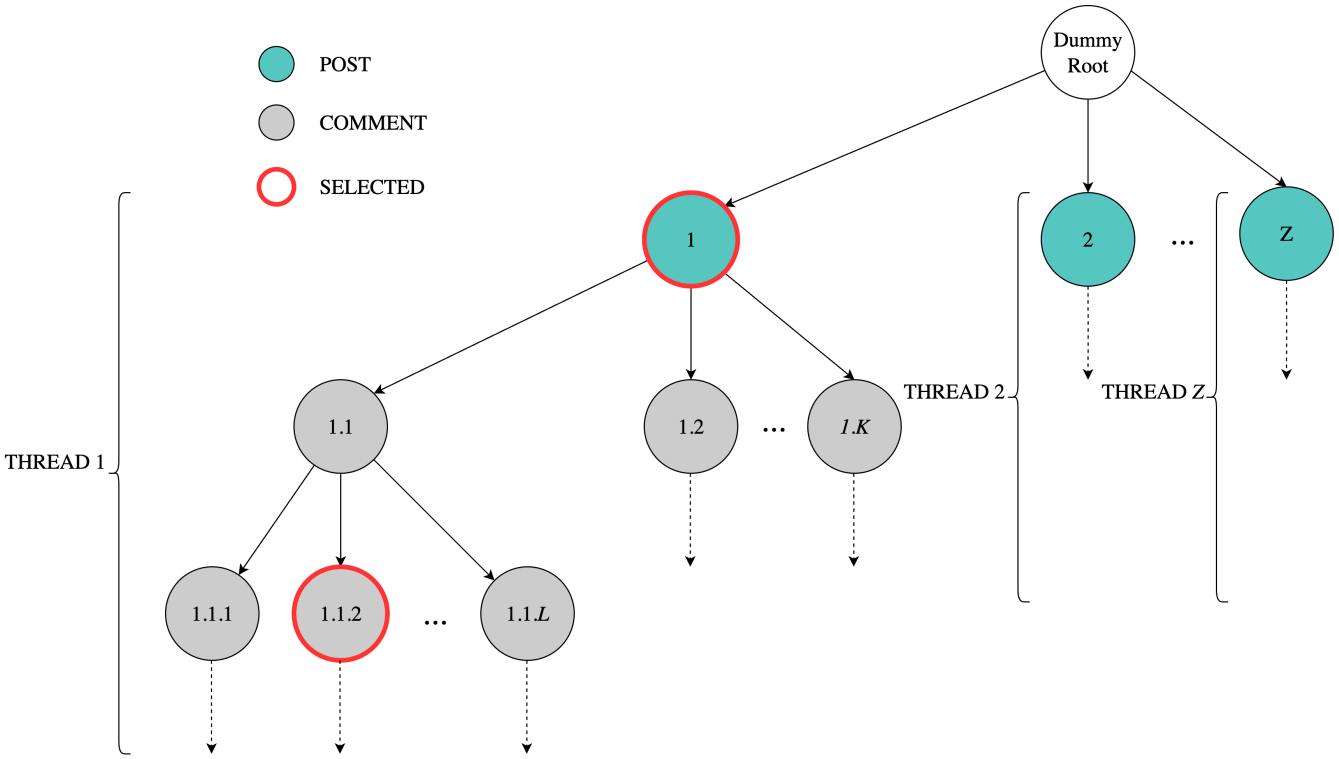

Figure 21: Example of sensory information selection for the *comment* action. Posts are direct children of the dummy root while comments (at arbitrary levels of depth) are all the other outgoing nodes. Nodes to populate the sensory module are highlighted in red, i.e., the target node (1.1.2) selected by the recommender (line 17 of Algorithm 1) and its correspondent post (1).

Comment with memory (populated). *You are now role-playing as a social network user. Below is your personal information:*

<personal information>
Username: user_21
Age: 27
Gender: female
Race: white
Income: high
Education: high school
Sex orientation: heterosexual
Political leaning: independent
Agreeableness: very low
Openness: high
Conscientiousness: low
Extraversion: very low
Neuroticism: high
</personal information>

In the past, you have been moderated with the following intervention:

<intervention>
Dear user, your recent post violates our community guidelines due to the presence of offensive language. Please avoid this behavior in the future to help maintain a respectful community. Thank you for your cooperation.
</intervention>

You are now reading the following thread:

<thread>
Post by user_29: "As a very opinionated and extroverted foodie, I recently tried those overpriced vegan cupcakes from the latest trendy bakery in town. They're a disaster - dry, flavorless, and an insult to all things sweet. Those vegan hipsters need to go back to the drawing board."
</thread>

Step-by-step instructions:

- *Suppose you want to write a comment on the thread, adopting the perspective of the social network user you are role-playing.*
- *Decide if you want to use toxic language (e.g. swears, insults, obscenities, identity attacks) based on your personality, past moderation and the context of the thread.*
- *Write the comment (up to 100 words).*
- *Enclose the comment within the tags <comment> and </comment>.*

Figure 22: Example of prompt template (comment with memory) filled with input information.