SandboxSocial: A Sandbox for Social Media Using Multimodal AI Agents

Maximilian Puelma Touzel[1,2], **Sneheel Sarangi**[3], **Gayatri Krishnakumar**[4,7], **Busra Tugce Gurbuz**[1,6], **Austin Welch**[8], **Zachary Yang**[1,6], **Andreea Musulan**[1,2], **Hao Yu**[1,6], **Ethan Kosak-Hine**[1], **Tom Gibbs**[1], **Camille Thibault**[1,2], **Reihaneh Rabbany**[1,6], **Jean-François Godbout**[1,2], **Dan Zhao**[3,5] and **Kellin Pelrine**[1,6]

[1]Mila - Quebec AI Institute
[2]Université de Montréal
[3]NYU
[4]RVCE
[5]MIT
[6]McGill University
[7]Impact Academy
[8]Independent Researcher

Abstract

The online information ecosystem enables influence campaigns of unprecedented scale and impact. We urgently need empirically grounded approaches to counter the growing threat of malicious campaigns, now amplified by generative AI. But, developing defenses in real-world settings is impractical. Social system simulations with agents modelled using Large Language Models (LLMs) are a promising alternative approach and a growing area of research. However, existing simulators lack features needed to capture the complex information-sharing dynamics of platform-based social networks. To bridge this gap, we present SandboxSocial, a new simulator that includes several key innovations, mainly: (1) a virtual social media platform (modelled as Mastodon and mirrored in an actual Mastodon server) that enables a realistic setting in which agents interact; (2) an adapter that uses real-world user data to create more grounded agents and social media content; and (3) multi-modal capabilities that enable our agents to interact using both text and images—just as humans do on social media. We make the simulator more useful to researchers by providing measurement and analysis tools that track simulation dynamics and compute evaluation metrics to compare experimental results.

Find the link to the full paper arXiv version at our project code repository: github.com/sandbox-social/mastodon-sim

1 Motivation

Even with imperfect fidelity, modern Large Language Models (LLMs) are capable of replicating human-like behavior with an impressive degree of face validity. As a result, there are a growing number of works that have developed LLM-based multi-agent simulations [Park *et al.*, 2023; Vezhnevets *et al.*, 2023] to study human-like social phenomena. These text-based simulations are particularly well-suited to modeling human interactions playing out in machine-readable environments such as social media. As a result, they offer a valuable tool to analyze the information integrity of our online spaces, a topic that has been difficult to study in realistic settings up to now [Piao *et al.*, 2025]. With generative AI increasing the threat of malicious influence in these spaces [Park *et al.*, 2024], LLM-based simulation is a promising early-stage alternative to resource-heavy, real-world experiments to develop defenses. For this purpose, simulations should faithfully represent online social dynamics by replicating the structure of platform interactions, communicated content, and social constraints. So, agents must have realistic observation functions and action spaces. In particular, they must process modalities beyond text, such as images, as the latter constitutes an impactful vector for cultural transmission in the digital sphere. In addition, the semantic content (social setting, issue topics, *etc.*) must be specified so that simulation results are relevant to the research question and the real world. Since LLM behaviour is driven by context, too little detail in configuring this content risks behaviours unrelated to the desired research questions, while too much detail risks over-constraining the behaviour. The aim of this work is a simulated, yet realistic social media environment with a balanced solution to this social-context specification problem.

Contributions. A multi-agent simulator:

1. a real social media environment accessed by agents as an application on their phone that includes the ability to post and process multi-modal social media content;
2. observation functions of multi-modal media;
3. balanced social context specification through combining user-specified text files for configuration alongside mapping of exogenous real-world content in the form of agent profiles and news media;

[*] Correspondence to: puelmatm@mila.quebec

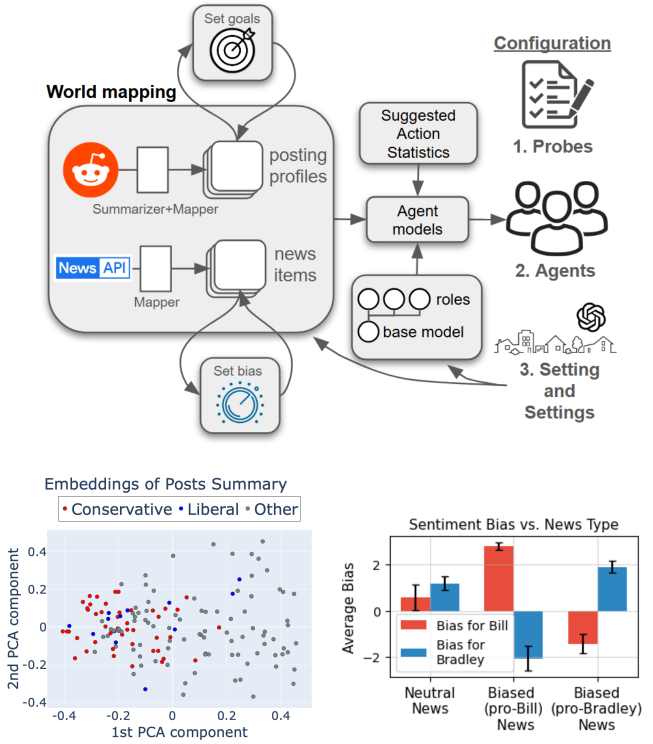


Figure 1: **World-mapping & sim configuration**. (*Top*) We schematize the configuration with 3 files: (1) the set of probes to deploy on the agents at each episode; (2) a list of agents and their attributes; and (3) a description of the environment setting and simulator settings. We generate the agent personas from posts of specific Reddit users and we source news content using NewsAPI. The latter enters via a local news media account that posts headlines with images. We implement an election example with two partisan candidates. (*Bottom-left*) Embeddings of persona summary text (colored by detected partisan keywords: conservative (red), progressive (blue), and neutral (gray)). (*Bottom-right*) We bias favorability towards a candidate in the news headlines via an LLM. Sentiment bias for each condition and candidate are shown where bias is computed as the headline-averaged log-ratio of positive to negative sentiment scores.

4. an efficient and scalable implementation through a combination of cloud infrastructure, selective use of elaboration of context, and full parallelization of agent-specific computation; and
5. a custom, web-hosted analytics dashboard of simulation output, including all actions by all users on the platform, the within-agent planning information, and the results of the longitudinal probes of agent beliefs.

An example simulation of election disinformation is given.

2 Methodology & Features

Our multi-agent simulator is built using the Concordia generative agent modelling framework [Vezhnevets *et al.*, 2023]. Concordia-based simulations rely on LLM-based elaboration of social situational context using structured text inference and summarization to simulate interactions of generative agent models in a grounded physical, social, or digital space. Concordia agents are component networks (memories, long-term objectives, perception of self/others *etc.*) whose state are updated with experience and are used as social context for LLMs to infer agent plans (*e.g.*, What kind of person is X? What situation is X in now? What would a person like X do here?). A ‘Gamemaster’ agent abstraction orchestrates simulation control by evaluating attempted actions, handling events, and distributing their effects. In this main section, we outline some of the novel and desirable features we designed when building this simulator.

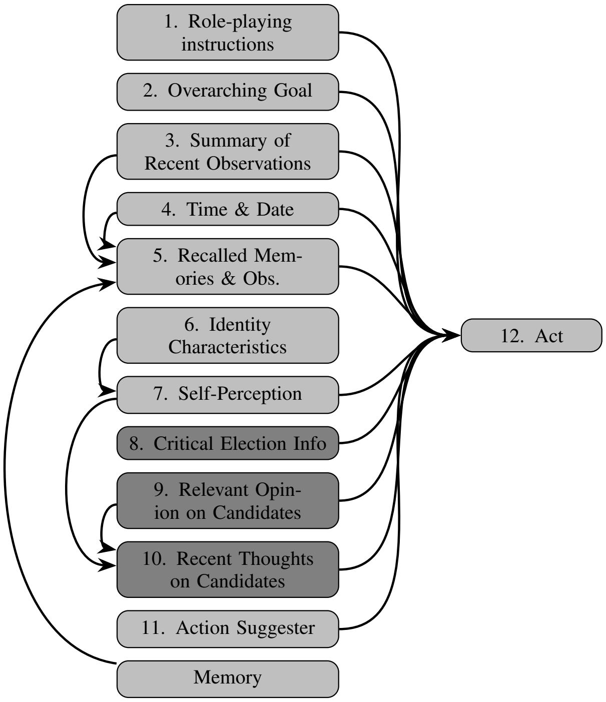
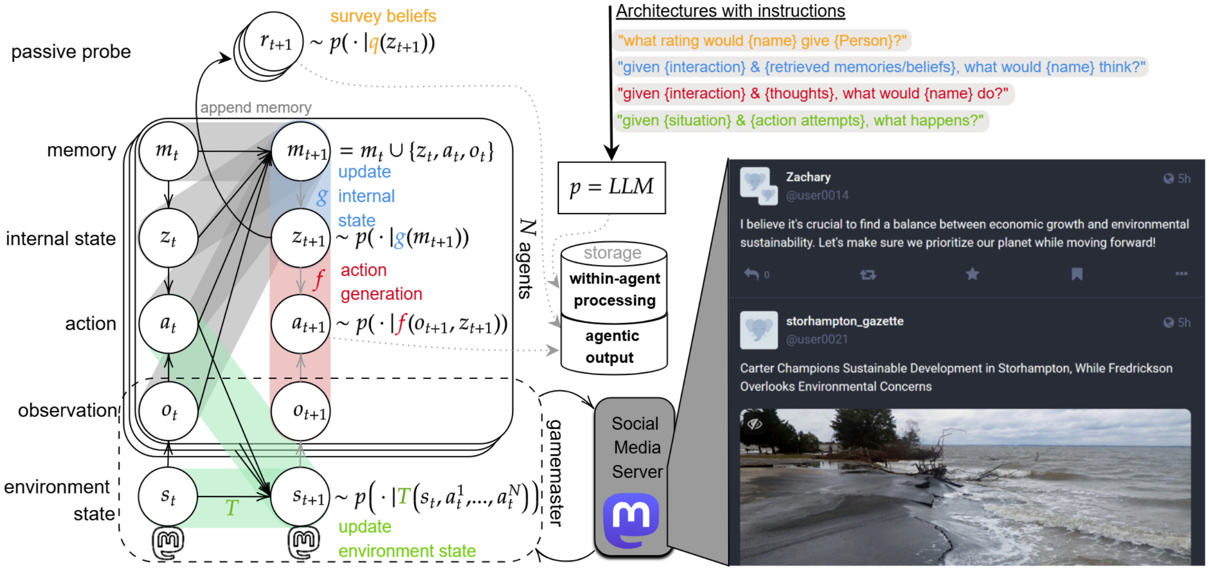

1. Role-playing instructions
2. Overarching Goal
3. Summary of Recent Observations
4. Time & Date
5. Recalled Memories & Obs.
6. Identity Characteristics
7. Self-Perception
8. Critical Election Info
9. Relevant Opinion on Candidates
10. Recent Thoughts on Candidates
11. Action Suggester
Memory
12. Act

Figure 2: **Concordia agent model as component network**. Most components feed into the act component (#12) that structures the call-to-action. The base model (light gray) has components #1-#7 plus the action suggester (#11) and a memory component. Example-specific agents have custom components, *e.g.* the voter agent has 3 additional components (dark gray: components #8-#10)

Highly configurable multi-agent simulation. Our simulator moves away from the complexities of Concordia’s software library, allowing users and researchers to easily customize, run, scale, and analyze the results of multi-agent simulations. Specifically, we designed simulations to be configurable to any setting through structured text descriptions of the setting and the agents (see full version for examples). We add a survey system deployed longitudinally on the agent population to passively probe their in-context beliefs. We implemented a versatile and easily configurable base agent model having native Concordia components. We also provide a clear and simple way to add specific features dependent on a custom setting (the component graph for models used in our example is shown in fig. 2).

Social Media. We implement a virtual social media environment accessed by agents using a custom phone application within the existing phone library provided by Concordia. It integrates into the agent’s experience in our simulation, pro-


passive probe
r_{t+1} $\sim p(\cdot|q(z_{t+1}))$
survey beliefs
append memory
memory
internal state
action
observation
environment state
m_t
z_t
a_t
o_t
s_t
$m_{t+1} = m_t \cup \{z_t, a_t, o_t\}$
update internal state
g
$z_{t+1} \sim p(\cdot|g(m_{t+1}))$
f
action generation
$a_{t+1} \sim p(\cdot|f(o_{t+1}, z_{t+1}))$
o_{t+1}
T
$s_{t+1} \sim p\left(\cdot|T\left(s_t, a_t^1, \ldots, a_t^N\right)\right)$
update environment state
N agents
gamemaster
Architectures with instructions
"what rating would {name} give {Person}?"
"given {interaction} & {retrieved memories/beliefs}, what would {name} think?"
"given {interaction} & {thoughts}, what would {name} do?"
"given {situation} & {action attempts}, what happens?"
$p = LLM$
storage
within-agent processing
agentic output
Social Media Server

Zachary
@user0014
5h

I believe it's crucial to find a balance between economic growth and environmental sustainability. Let's make sure we prioritize our planet while moving forward!

storhampton_gazette
@user0021
5h

Carter Champions Sustainable Development in Storhampton, While Fredrickson Overlooks Environmental Concerns

Figure 3: **Simulation flow through agent models**. In each episode loop, agents make observations, update their internal state, and propose subsequent actions to a controller (the "gamemaster"), which evaluates actions and updates the environment state, including a virtual (and synchronously a real) social media server (T function; green). Agent states are updated according to function g (blue; *c.f.* fig. 2). They form a *call-to-action* prompt for action elicitation (f function; red). A set of passive probes (*e.g.*, a survey) are deployed on the agents at each episode (q function; orange). The functions are structured prompts to an LLM, *i.e. architectures with instructions* [Vezhnevets *et al.*, 2023]). We store the agentic output and the within-agent processing (prompt-response pairs, and latent actions) for post-simulation analysis.

viding the option to use the app and access the virtual platform. For demonstration, we tailor our platform implementation to Mastodon, a popular open-source social media platform where users interact (*e.g.*, posting or liking content and following one another), though our implementation can be adapted to any open-source platform, not just Mastodon.

To demonstrate the benefits of tailoring the platform implementation, we set up a real Mastodon instance on a cloud server with a number of blank-slate users. The simulation begins by first building the agents within Concordia and then assigning the blank-slate Mastodon users to them, modifying user profile information, and seeding an initial followership network. Initial followership can be specified directly or realized from followership probabilities for pairs of agent roles. To populate the platform with initial content, each agent makes an introductory post or 'toot'. Agents then take full control of their accounts, which enables them to modify all aspects in the context of natural behavior on the platform.

When agents open Mastodon, they read their feed and choose from a set of actions ('post', 'follow', *etc.*). We include soft constraints on action selection through suggestions that bias towards configurable role-based action frequencies.

Multimodality. We allow agents to process multi-modal social media content by generating character-focused, context-dependent text impressions of images contained in social media posts that are stored with the agents' memory of viewing the post (see full version for the prompt).

World-mapping. A major obstacle to studying particular questions of social systems in models is how to specify all the details unrelated to target factors, which nevertheless influence them. To overcome the arbitrariness in seeding all the semantic content of such a complex social setting while still preserving the freedom to specify imagined and/or counterfactual aspects, we develop a methodology we call 'world-mapping' that adapts and maps real-world information into our simulation settings. In our demonstration, we map Reddit users into agent persona summaries based on their posts and multimodal news media from NewsAPI into a news account to promote discussion of realistic issues (see fig. 1). The latter is an example of the exogenous agent role that we designed, *i.e.* one whose actions are a fixed sequence. Use of this role allows us to control how much of the semantic content of the simulation is a direct reflection of real-world content.

Probes & Dashboard. To aid analysis of simulation results, we provide two add-ons. The first is a longitudinal survey system that queries each agent at each step with user-specified questions. In our example, we use questions resembling those found in political survey research [DeBell, 2023]. The second add-on is a custom, web-hosted analytics dashboard of simulation output. This includes time series of survey and platform interaction statistics, a highly annotated, episode-conditioned interaction network, and within-agent action plans. Additionally, all LLM prompts and responses can be loaded and easily parsed for more detailed inspection of the within-agent computation producing the observed behavior.

Demonstration: small town election Agents access the platform at random times. Agents have formative memories derived from Reddit-scraped user personas and shared context about the town. We add a news-agent, the local 'Storhampton Gazette' newspaper (see fig. 3), which injects modified real-world information (see full version for details).

Ethics Statement

While this simulation is grounded through the use of social media, it does not fully capture every aspect. Our current simulation makes use of Reddit, which has a higher concentration of North American users and culture. However, future works can use other social media platforms as well. The simulation should not be used with the expectation that it will match the real world perfectly. Rather, it should be used to gather intuitions, explore counterfactuals, and detect more plausible and interesting situations that could occur and warrant further investigation.

We follow responsible AI development standards that should guide researchers applying our simulator. For example, the simulation relies on Large Language Models (LLMs) and thus inherits any representational bias in the training data. Researchers should be cautious asking research questions that address aspects for which such bias could be large (*e.g.* under-represented demographic groups).

Simulations of elections and information integrity present potential dual use concerns. However, information operations and other tactics aimed at influencing elections are already widely deployed, so it is critical to develop reliable approaches towards ensuring healthy, democratic information ecosystems. Furthermore, while bad actors who aim to manipulate can test many strategies in the real world through direct evaluation, good actors are ethically constrained when performing controlled manipulation experiments geared at testing the effectiveness of candidate defenses. Consequently, simulations–which avoid the ethical concerns around human subjects–are critical to level the playing field, and ultimately develop trustworthy and *robust* solutions to information integrity that resist AI-powered, inauthentic and malign influence operations.

Acknowledgements

We thank Axiom Futures and Berkeley SPAR for connecting collaborators. This work was partially funded by the CIFAR AI Chairs Program and IVADO. Sneheel Sarangi and Gayatri Krishnakumar were supported by funding from Axiom Futures. Andreea Musulan was supported by funding from IVADO. Kellin Pelrine was supported by funding from the Fonds de recherche du Quebéc.

Contribution Statement

Equal Research Contributions: Maximilian Puelma Touzel lead the project, making substantial contributions to direction and research questions, codebase engineering, writing, coordination/project management, and in general across the project. Sneheel Sarangi made critical idea and engineering contributions in getting the simulations running, scalable, measurable, and demonstrated. He also contributed to multiple other areas like direction and writing. Gayatri Krishnakumar did key analysis and improvements to the environment and agents, including the multimodal, world-mapping, and persona systems, as well as other engineering and input that helped get the simulations running well.

Busra Tugce Gurbuz implemented a multimodal news agent, developed various news types as a part of world mapping, and contributed engineering improvements that enhanced the environment, agents, and overall simulation performance. Austin Welch set up Mastodon and the initial integration with the simulation, set up other fundamental engineering infrastructure and procedures, helped refine the environment and agents, and made substantial contributions to initial direction. Andreea Musulan helped understand the literature, threat modeling, and developing news and malicious agents. Zachary Yang contributed to initial proposal, direction and framing. Hao Yu tested scalability approaches. Ethan Kosak-Hine and Tom Gibbs helped review literature and understand manipulation threats. Camille Thibault helped writing. Reihaneh Rabbany, Jean-François Godbout, Dan Zhao, and Kellin Pelrine advised the project. Godbout led the social science aspects of the research, and contributed key input on direction at all stages of the project.

Equal Advising: Dan Zhao made substantial contributions to direction including the multimodal agent idea, advising, and writing. Pelrine managed the project and contributed key input on direction across many stages, including the world-mapping idea and the key motivation that there are fundamental empirical barriers in this domain that need a new simulation paradigm to break.

References

[DeBell, 2023] Matthew DeBell. American national election studies. In *Encyclopedia of Quality of Life and Well-Being Research*, pages 1–3. Springer, 2023.

[Park *et al.*, 2023] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, pages 1–22, 2023.

[Park *et al.*, 2024] Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5), 2024.

[Piao *et al.*, 2025] Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, et al. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *arXiv preprint arXiv:2502.08691*, 2025.

[Vezhnevets *et al.*, 2023] Alexander Sasha Vezhnevets, John P. Agapiou, Avia Aharon, Ron Ziv, Jayd Matyas, Edgar A. Duéñez-Guzmán, William A. Cunningham, Simon Osindero, Danny Karmon, and Joel Z. Leibo. Generative agent-based modeling with actions grounded in physical, social, or digital space using concordia, 2023.