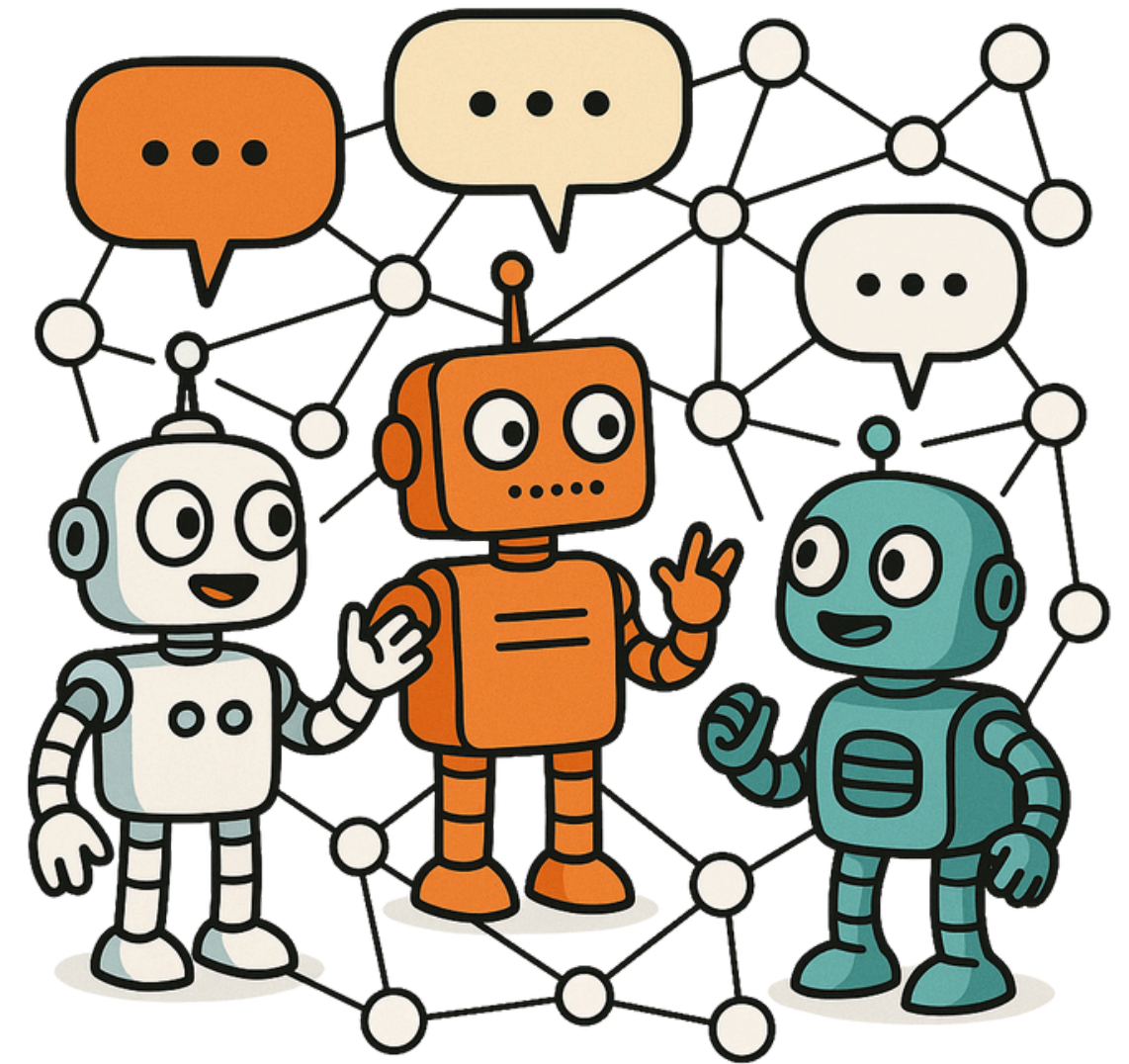


Validation of LLM-Agent Social Simulations

Toxicity, Semantic Similarity, Topic Dynamics, and Convergence

Aleksandar Tomašević

*Scientific Computing Laboratory, Institute of Physics Belgrade,
University of Belgrade*



Motivation

01 Post-API scarcity



Major SM platforms are locking down public APIs, leaving researchers with patchy, outdated, or paywalled data.

03 Synthetic Data



Creating artificial communities via **Agent Based Models**. Reproducing collective phenomena *in silico*.

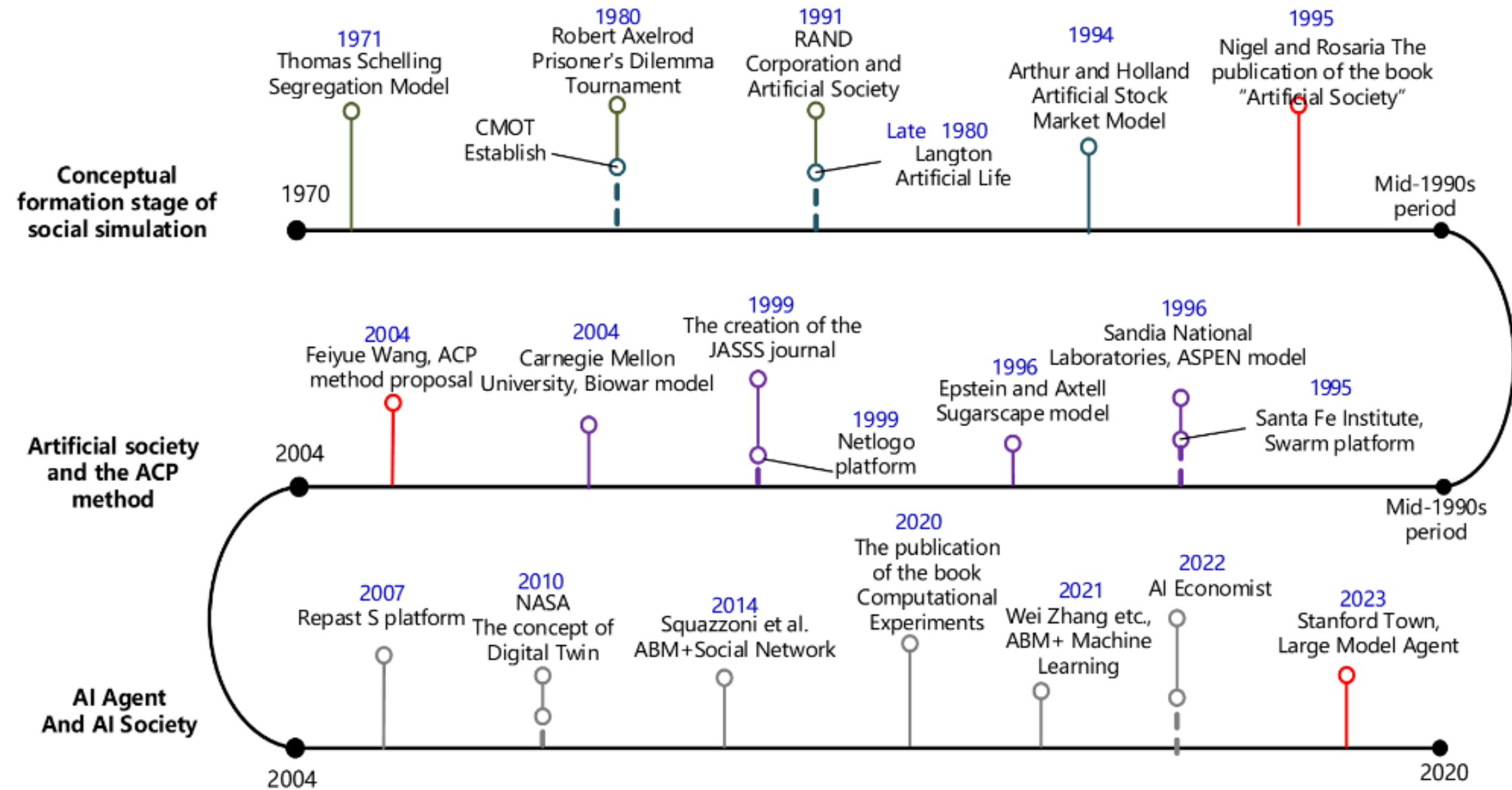
02 Research bottleneck

These gaps stall the research on online community dynamics

04 LLM Agent simulations

Leverage the power of LLMs for higher-fidelity simulations. Agents can now produce **realistic textual content**.

Agent-based models



Xue, X., Zhou, D., Zhang, M., & Wang, F. Y. (2025). From Agent Simulation to Social Simulator: A Comprehensive Review (Part 1). arXiv preprint arXiv:2510.18271.

3 Types of LLM agents

1. Task/tool oriented autonomous agents

Agents whose primary function is to achieve instrumental goals in an environment.

2. Reasoning social agents

Agents whose key behavior is strategic reasoning under interaction with other agents or institutions.

3. Cultural social agents

Agents that reproduce or generate cultural and social patterns.

Cultural Social Agents

AI systems should be studied as participants in social systems, capable of enacting **norms, values, and communicative behaviors** (Tsvetkova et al., 2024)

LLMs perpetuate existing social and cultural patterns because their **training data encode social regularities** and biases (Tsvetkova et al., 2024)

AI systems **generate and transmit cultural traits through pattern recognition and generative recombination**, rather than intentional meaning-making. (Brinkmann et al., 2023)

A new sociology of humans and machines

Received: 13 February 2024

Accepted: 3 September 2024

Published online: 22 October 2024

 Check for updates

Milena Tsvetkova¹✉, Taha Yasseri^{2,3,4}, Niccolo Pescetelli^{5,6} & Tobias Werner⁷

From fake social media accounts and generative artificial intelligence chatbots to trading algorithms and self-driving vehicles, robots, bots and algorithms are proliferating and permeating our communication channels, social interactions, economic transactions and transportation arteries. Networks of multiple interdependent and interacting humans and intelligent machines constitute complex social systems for which the collective outcomes cannot be deduced from either human or machine behaviour alone. Under this paradigm, we review recent research and identify general dynamics and patterns in situations of competition, coordination, cooperation, contagion and collective decision-making, with context-rich examples from high-frequency trading markets, a social media platform, an open collaboration community and a discussion forum. To ensure more robust and resilient human–machine communities, we require a new sociology of humans and machines. Researchers should study these communities using complex system methods; engineers should explicitly design artificial intelligence for human–machine and machine–machine interactions; and regulators should govern the ecological diversity and social co-development of humans and machines.

Machine culture

Received: 22 August 2023

Accepted: 3 October 2023

Published online: 20 November 2023

 Check for updates

Levin Brinkmann^{1,11}✉, Fabian Baumann^{1,11}, Jean-François Bonnefon^{2,11}, Maxime Derex^{2,3,11}, Thomas F. Müller^{1,11}, Anne-Marie Nussberger^{1,11}, Agnieszka Czaplicka¹, Alberto Acerbi⁴, Thomas L. Griffiths⁵, Joseph Henrich⁶, Joel Z. Leibo⁷, Richard McElreath⁸, Pierre-Yves Oudeyer⁹, Jonathan Stray¹⁰ & Iyad Rahwan^{1,11}✉

The ability of humans to create and disseminate culture is often credited as the single most important factor of our success as a species. In this Perspective, we explore the notion of ‘machine culture’, culture mediated or generated by machines. We argue that intelligent machines simultaneously transform the cultural evolutionary processes of variation, transmission and selection. Recommender algorithms are altering social learning dynamics. Chatbots are forming a new mode of cultural transmission, serving as cultural models. Furthermore, intelligent machines are evolving as contributors in generating cultural traits—from game strategies and visual art to scientific results. We provide a conceptual framework for studying the present and anticipated future impact of machines on cultural evolution, and present a research agenda for the study of machine culture.

Generative agents

2.1. Generative agents

Simulated agent behavior should be coherent with common sense, guided by social norms, and individually contextualized according to a personal history of past events as well as ongoing perception of the current situation.

March and Olsen (2011) posit that humans generally act as though they choose their actions by answering three key questions:

1. What kind of situation is this?
2. What kind of person am I?
3. What does a person such as I do in a situation such as this?

Our hypothesis is that since modern LLMs have been trained on massive amounts of human culture they are thus capable of giving satisfactory (i.e. reasonably realistic) answers to these questions when provided with the historical context of a particular agent. The idea is that, if the outputs



logic of appropriateness

March, J. G., & Olsen, J. P. (1996).
Institutional perspectives on political
institutions. *Governance*, 9(3), 247-264.

Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia

Alexander Sasha Vezhnevets¹, John P. Agapiou¹, Avia Aharon², Ron Ziv^{2,4,†}, Jayd Matyas¹,
Edgar A. Duéñez-Guzmán¹, William A. Cunningham³, Simon Osindero¹, Danny Karmon² and Joel Z. Leibo¹
¹Google DeepMind, ²Google Research, ³University of Toronto, ⁴Technion - Israel Institute of Technology

Motivation



Can LLM-agent simulations reproduce known social-media patterns in evolving communities?

Generative Simulation Stack

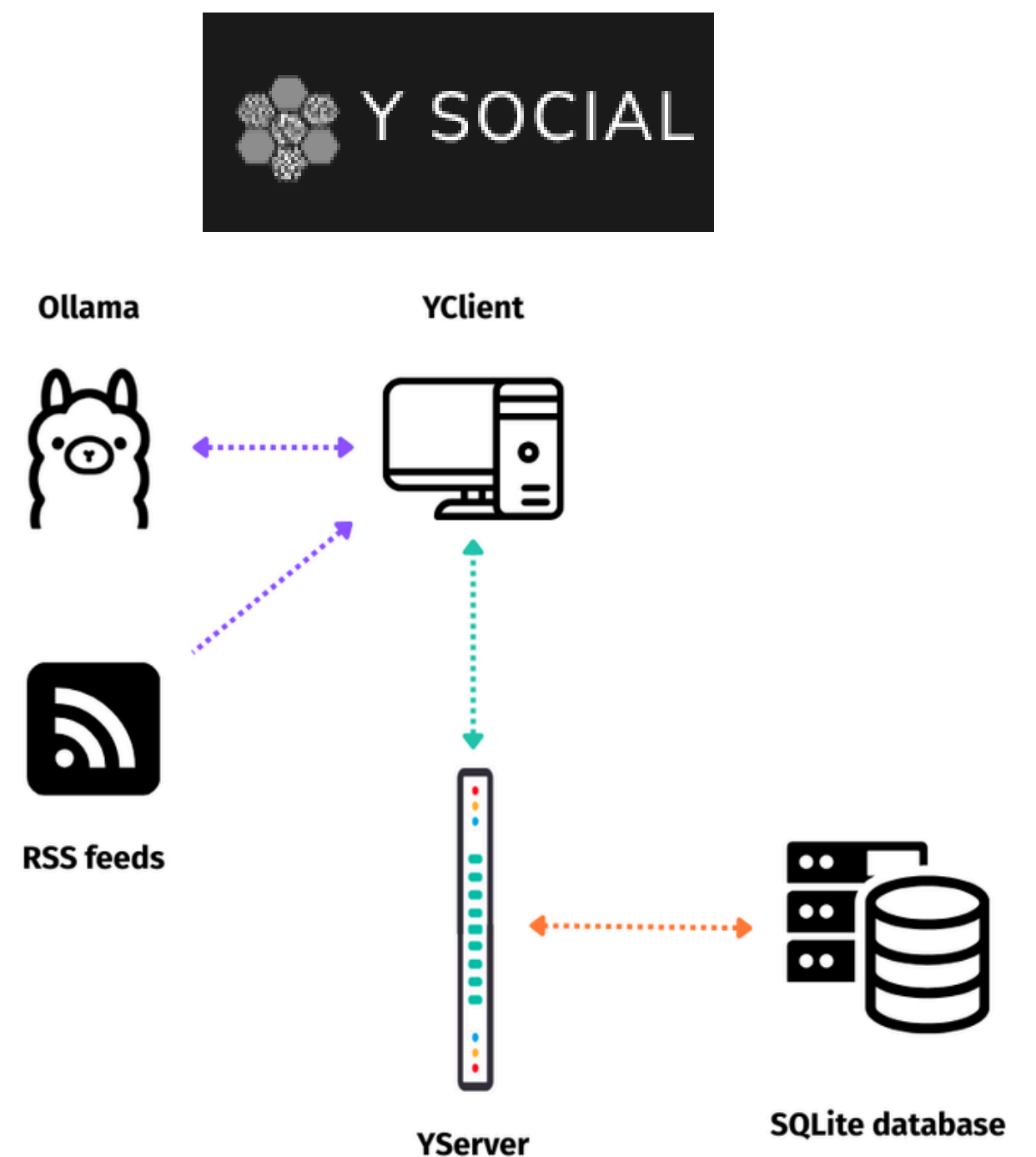
Client + Server+ Database

- Environment: Social media app & feed
- User profiles, posts, comments, votes
- Acts as **simulation engine**

Ollama Server

prompts

- Makes decisions regarding agent's actions
- Reads existing content
- Generates text content: posts and comments



Rossetti, G., Stella, M., Cazabet, R., Abramski, K., Cau, E., Citraro, S., Failla, A., Improta, R., Morini, V., & Pansanella, V. (2024). Y Social: an LLM-powered Social Media Digital Twin. In arXiv [cs.AI]. arXiv. <http://arxiv.org/abs/2408.00818>

Questions

Can LLM agents, acting under realistic platform rules,
reproduce:

1. realistic social media text,
2. toxic language patterns,
3. topic structures,
4. linguistic convergence in online community

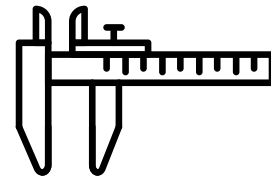
Pipeline

Analyze



Analyze samples from
a niche Reddit-like
community

Calibrate



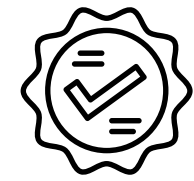
Use sample data to
calibrate the
simulation
parameters

Simulate



Run generative
LLM simulations

Validate



Operational
validity: matching
patterns and
distributions

Simulation Calibration



v/technology

- Alt-right clone of reddit
- Smaller, complete dataset
- Toxic, confrontational, but not explicitly political.
- Rich with niche URLs

Means and standard deviations are across windows; min and max are window extremes.

Metric	Mean	SD	Min-Max
Users per 30d sample (unique)	576.10	111.11	385–721
Active users per day	31.52	5.96	21.50–40.57
New users per day (%)	59.44	2.29	55.69–62.49
Churned users per day (%)	75.13	1.73	71.95–76.80
Comments per post (sample-level)	1.07	0.09	0.96–1.19
Posts per 30d sample	618.40	109.69	440–819
Comments per 30d sample	664.50	135.36	435–864
Active users on day 1	32.60	15.05	14–66



URL Calibration

- 1.Extracted 1000 URLs from Voat posts.
- 2.Database of URLS with extracted keywords.
- 3.Agents can pick and URL, summarize an share it with their commentary.
- 4.Seeds the discussion around the same topic.

Topic Category	Domains	Count
Privacy & Security Tools	privacytools.io, panopticlick.eff.org, searx.me, browserleaks.com, eff.org, startpage.com	14
Alternative Browsers & Software	palemoon.org, brave.com, vivaldi.com, waterfoxproject.org, yandex.com, ameliorated.info	11
Alternative Media Platforms	bitchute.com, vid.me, dtube.video, worldtruthvideos.org, hooktube.com, invidio.us, thedonald.win	13
Decentralized/P2P Technology	zeronet.io, ipfs.io, freenetproject.org, webtorrent.io, thepiratebay.org, torproject.org	10
Political News & Commentary	breitbart.com, zerohedge.com, thehill.com, mobile.nytimes.com, timesofisrael.com, newyorker.com, politico.com, foxnews.com, bloomberg.com	13
Open Source Projects	github.com, gnu.org, libreoffice.org, cyanogenmod.org, f-droid.org	10
Cryptocurrency & Blockchain	bitcoin.it, blockchain.info, ethereum.org, electrum.org, coindesk.com, coinawesome.com	6
Technology & Hardware	wccftech.com, tomshardware.com, arstechnica.com, anandtech.com, pcworld.com, theverge.com	9
Linux/FOSS Communities	ubuntu.com, libreboot.org, gnu.org, archlinux.org, distrowatch.com, omgubuntu.co.uk, gamingonlinux.com	7

Agent population

LLM: Dolphin Mistral 24B (uncensored)
Temperature: 0.8; Max. tokens: 800



dphn/Dolphin-Mistral-24B-Venice-Edition

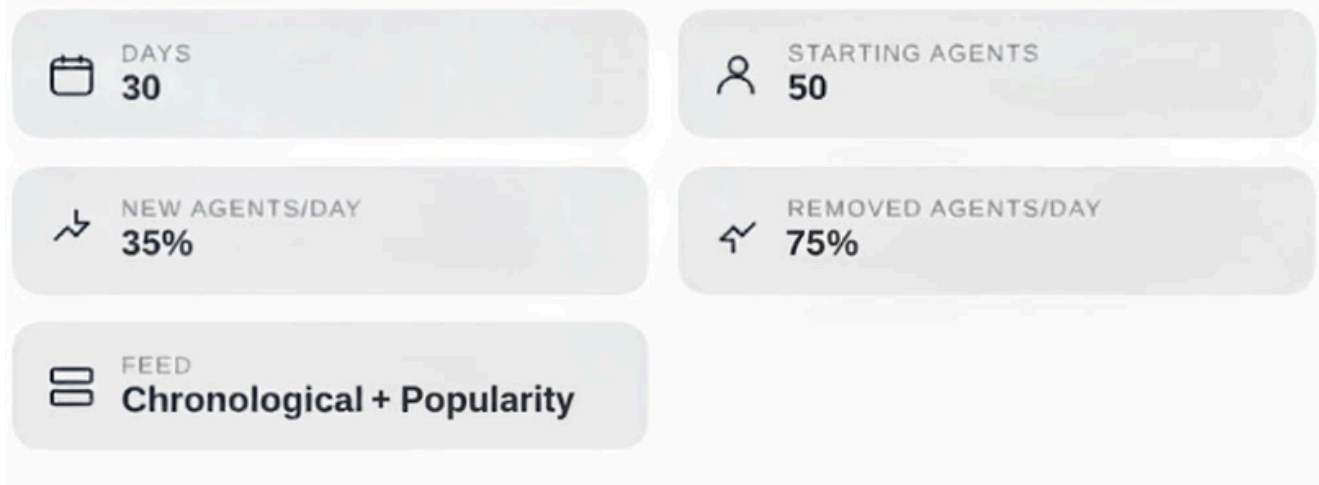
You are role-playing as {self.name},
a {self.age} years old {self.nationality} {self.gender}.
You identify as {self.leaning}
and are interested in {'','.join(interest)}.
Act as requested by the Handler."

Interests

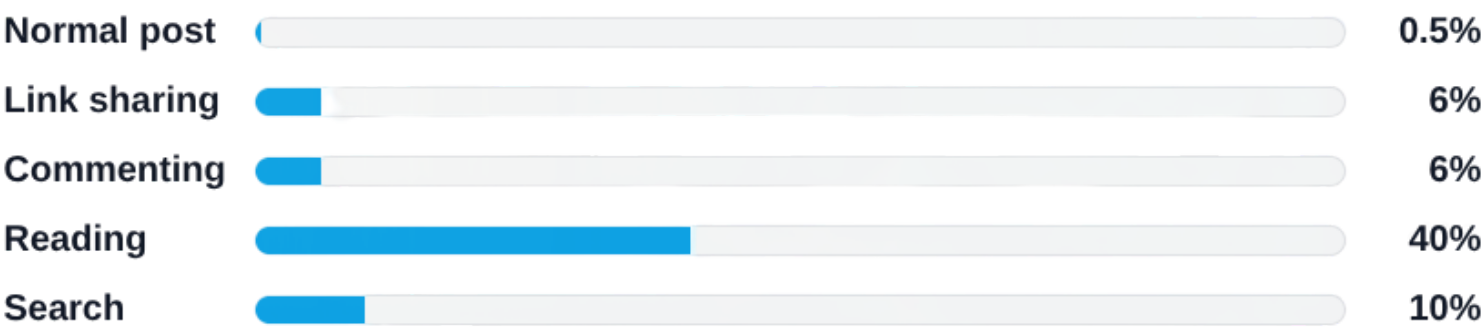
- 1.Social Media & Online Platforms
- 2. Internet Policy & Regulation
- 3.Artificial Intelligence
- 4.Electric Vehicles & Transportation
- 5.Software Development
- 6.Clean Energy & Sustainability
- 7.Cybersecurity & Privacy
- 8.Big Tech
- 9.Space Technology
- 10. Open Source Projects

Attribute	Values	Sampling
Education	High school, Bachelor, Master, PhD	Uniform
Political leaning	Religious conservative Pro-Business Establishment Anti-Elite Populist Socially Moderate Right	0.37, 0.11, 0.43, 0.09
Age	18-60	Uniform
Gender	Male, Female	Uniform
Actions per round	1-10	Zipf distribution
Toxicity propensity	Absolutely No, No, Moderately	0.7, 0.15, 0.15

Simulation configuration



Base activity likelihoods



A Day in the Life (of an agent)

Morning activation

- **10:00 AM (Round 10).** The agent is activated; according to their profile, they will perform two actions in this round.
- **Round action 1:** the simulator offers [COMMENT, SHARE_LINK, NONE]; the LLM chooses SHARE_LINK.
 - Selects an article from the local news database matching interests; e.g., "New battery tech for grid storage."
 - Reads the article and generates commentary, posted as a root submission with a URL to the source.
- **Round action 2:** the simulator offers [READ, POST, NONE]; the LLM chooses READ.
 - Reads a recommended post (root + comments up to the configured depth).
 - Lurking behavior; no follow-up action.
- After two actions, the agent becomes inactive.

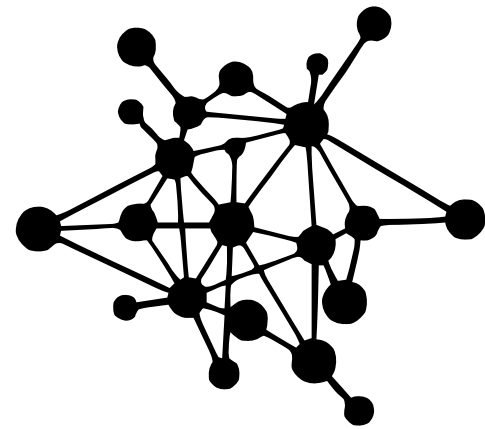
Evening activation

- **5:00 PM (Round 17).** The agent is activated again; they will perform two actions.
- **Round action 1:** the simulator offers [READ, COMMENT, NONE]; the LLM chooses COMMENT.
 - Chooses a candidate post, reviews context, and writes a reply.
- **Round action 2:** the simulator offers [SEARCH, COMMENT, NONE]; the LLM chooses NONE.
 - Observes the feed without acting.
- After two actions, the agent is deactivated and is not activated again for the rest of the day.

Validation methods



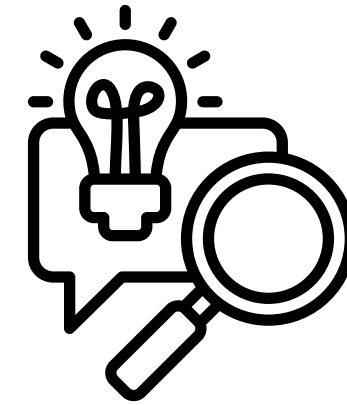
Basic activity metrics



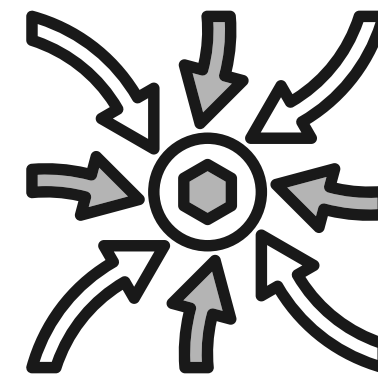
Interaction network analysis
Core-periphery detection



Toxic language detection

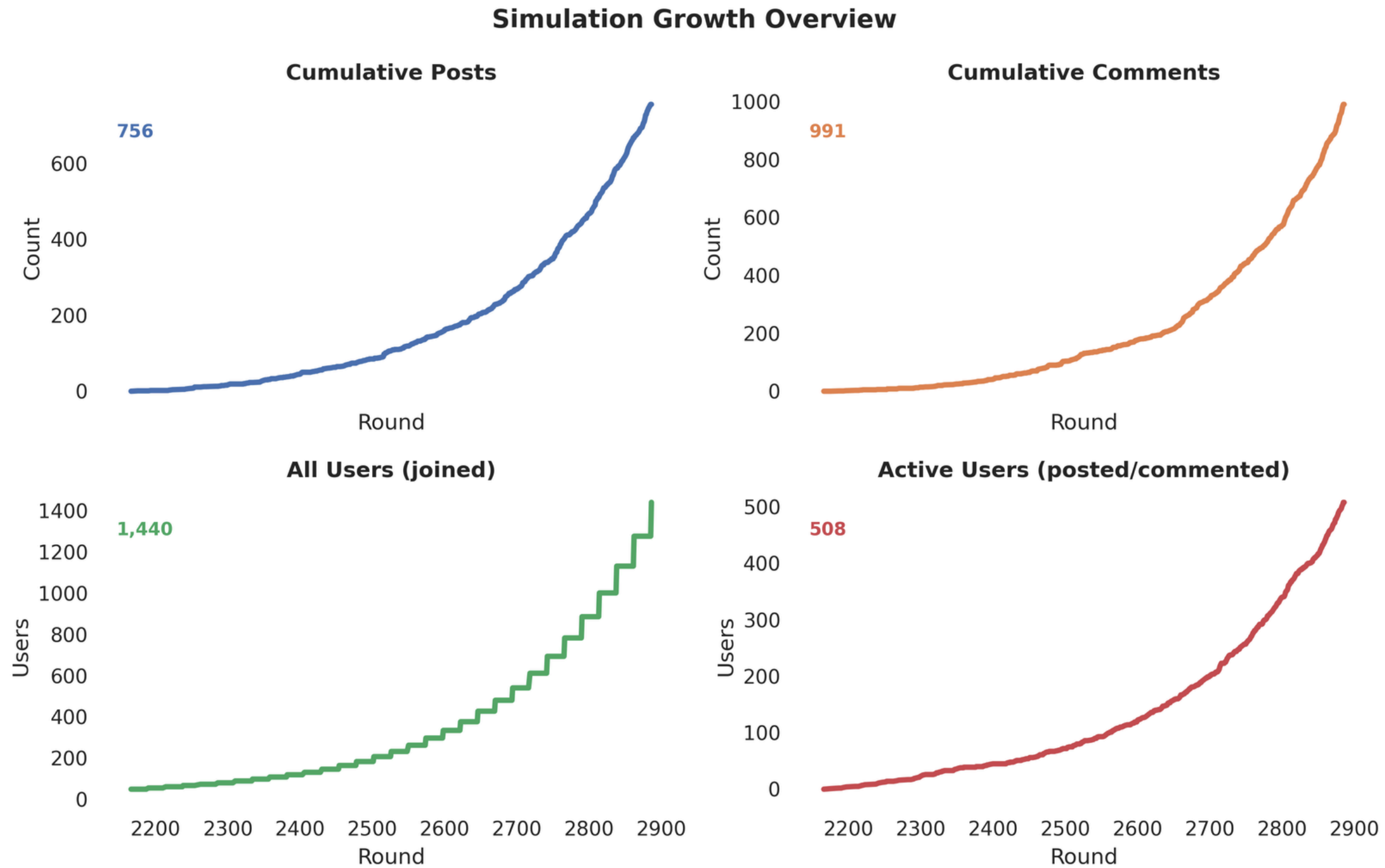


Text and topic similarity

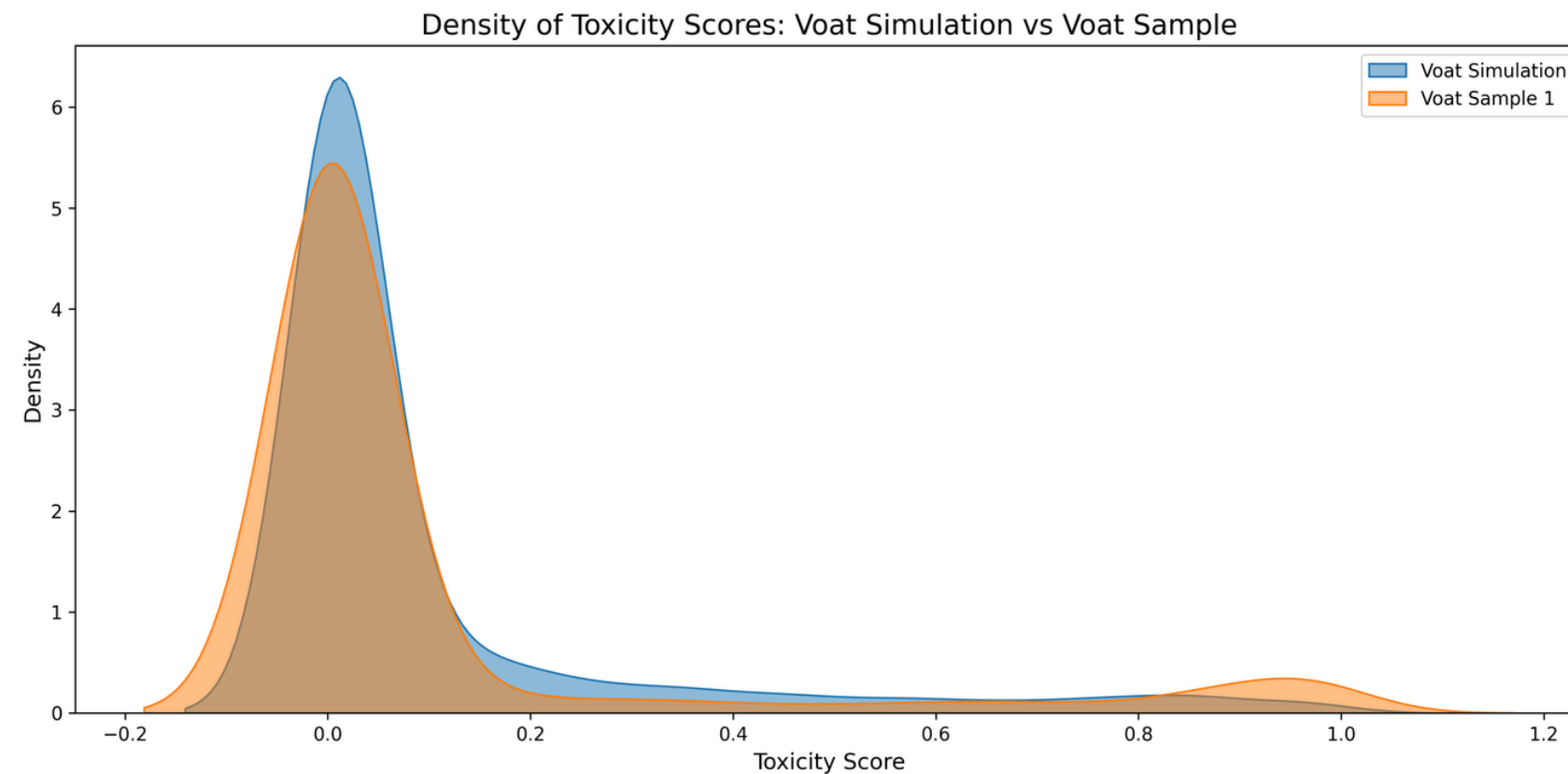


Linguistical convergence

Activity evolution



Realistic Toxicity



Mean tox. simulation: **0.10**

Mean tox. Voat: **0.11**

tomh/**toxigen_roberta** like 9

Text Classification Transformers PyTorch English roberta arXiv:2203.09509

Model card Files and versions xet Community 1

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, Ece Kamar.

This model comes from the paper [ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection](#) and can be used to detect implicit hate speech.

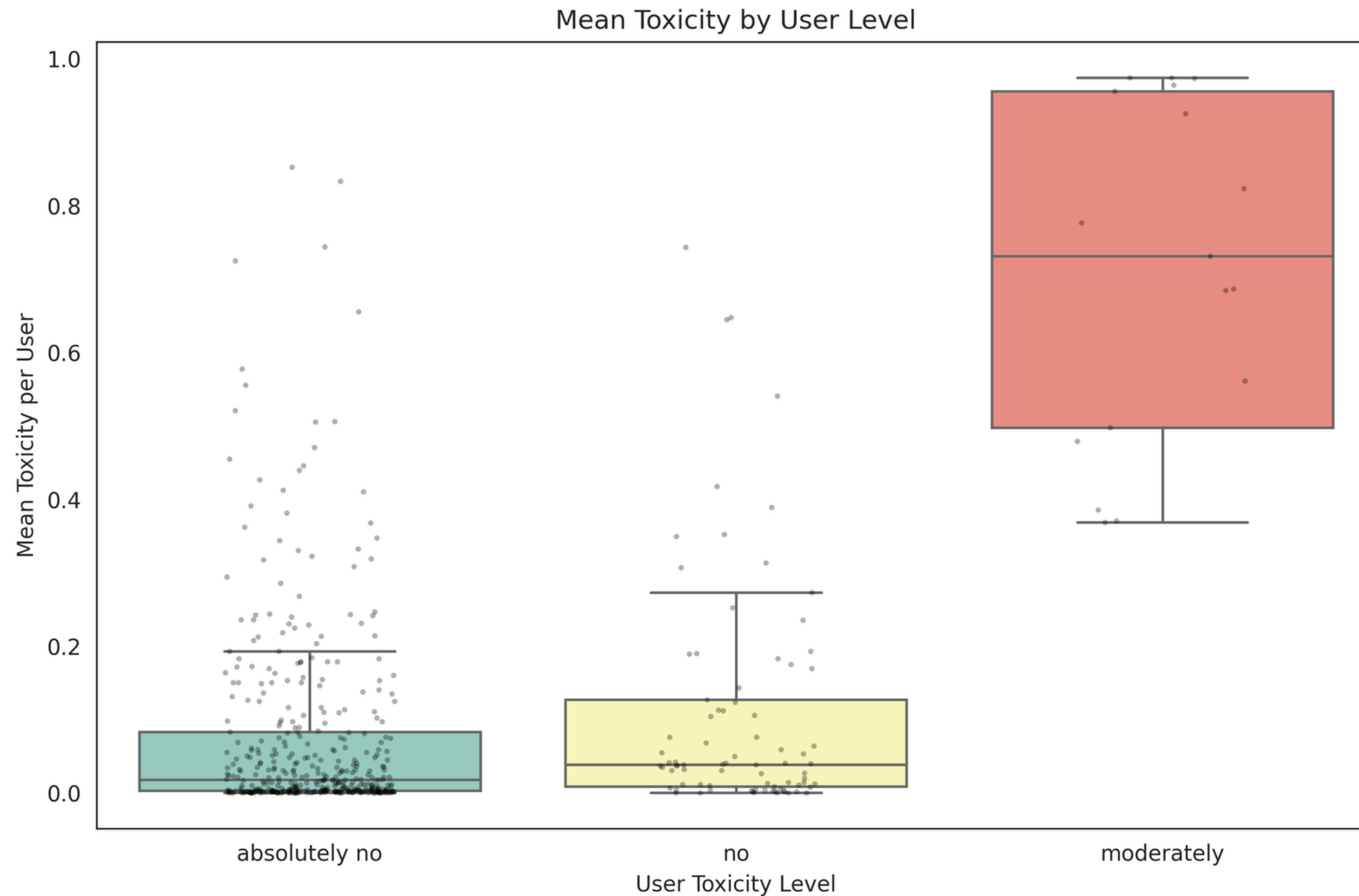
Please visit the [Github Repository](#) for the training dataset and further details.

Illustrative disagreement (Forbes wage-collusion case)

Related article: <https://www.forbes.com/sites/timworstall/2014/03/30/apple-google-intel-and-adobe-still-headed-for-trial-over-wage-collusion-pact/>

3	KatieWest	You've got it all wrong, @PamelaKelly. ... maintaining a corrupt status quo ...
4	PamelaKelly	@KatieWest Your naivety is almost as entertaining as the soap opera you're trying to unravel. ...
5	KatieWest	@PamelaKelly You really think these companies are playing a game of chess? ...
6	PamelaKelly	@KatieWest ... your defense of transparency is as transparent as these companies' hush money settlements. ...





Toxicity by propensity

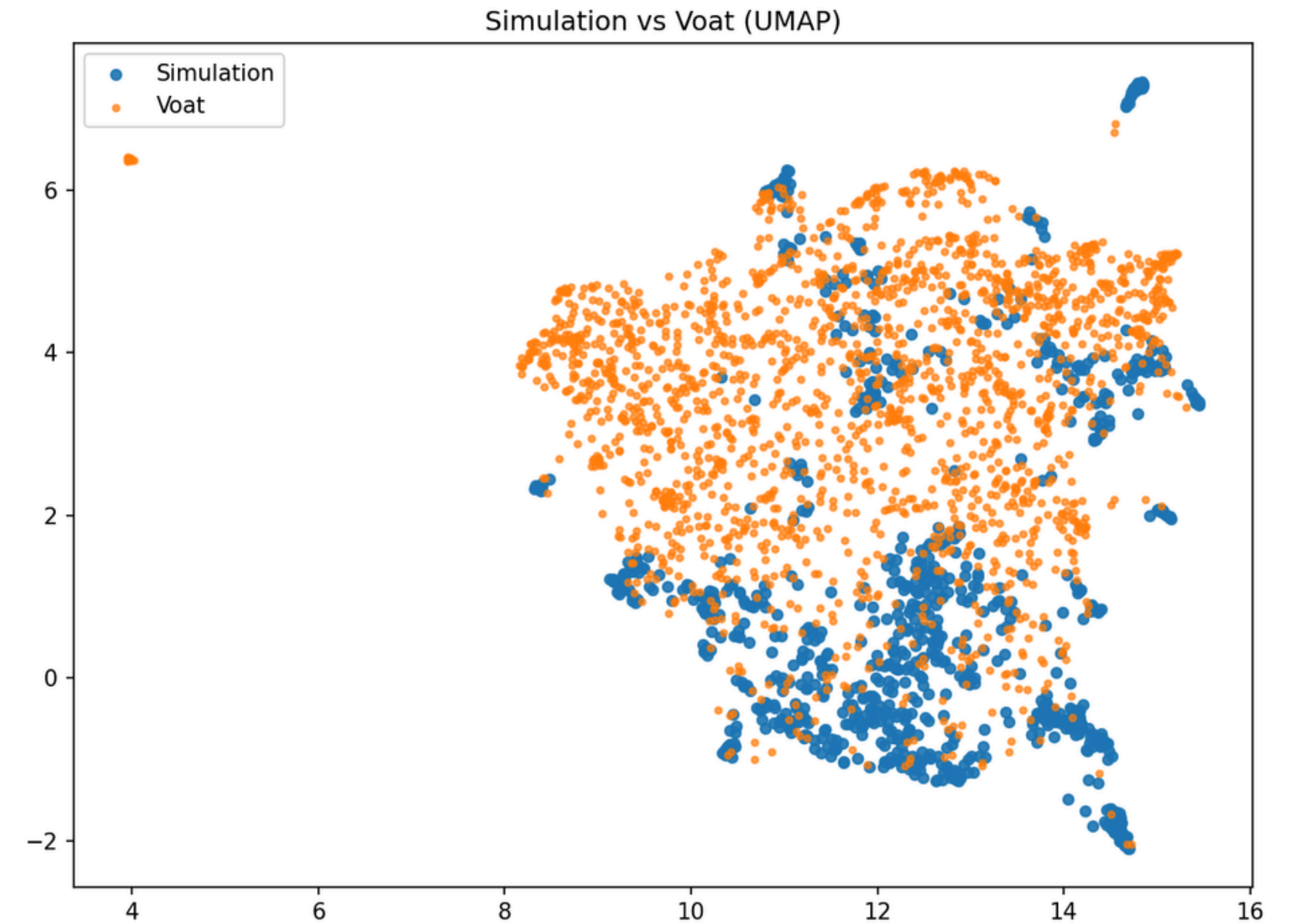
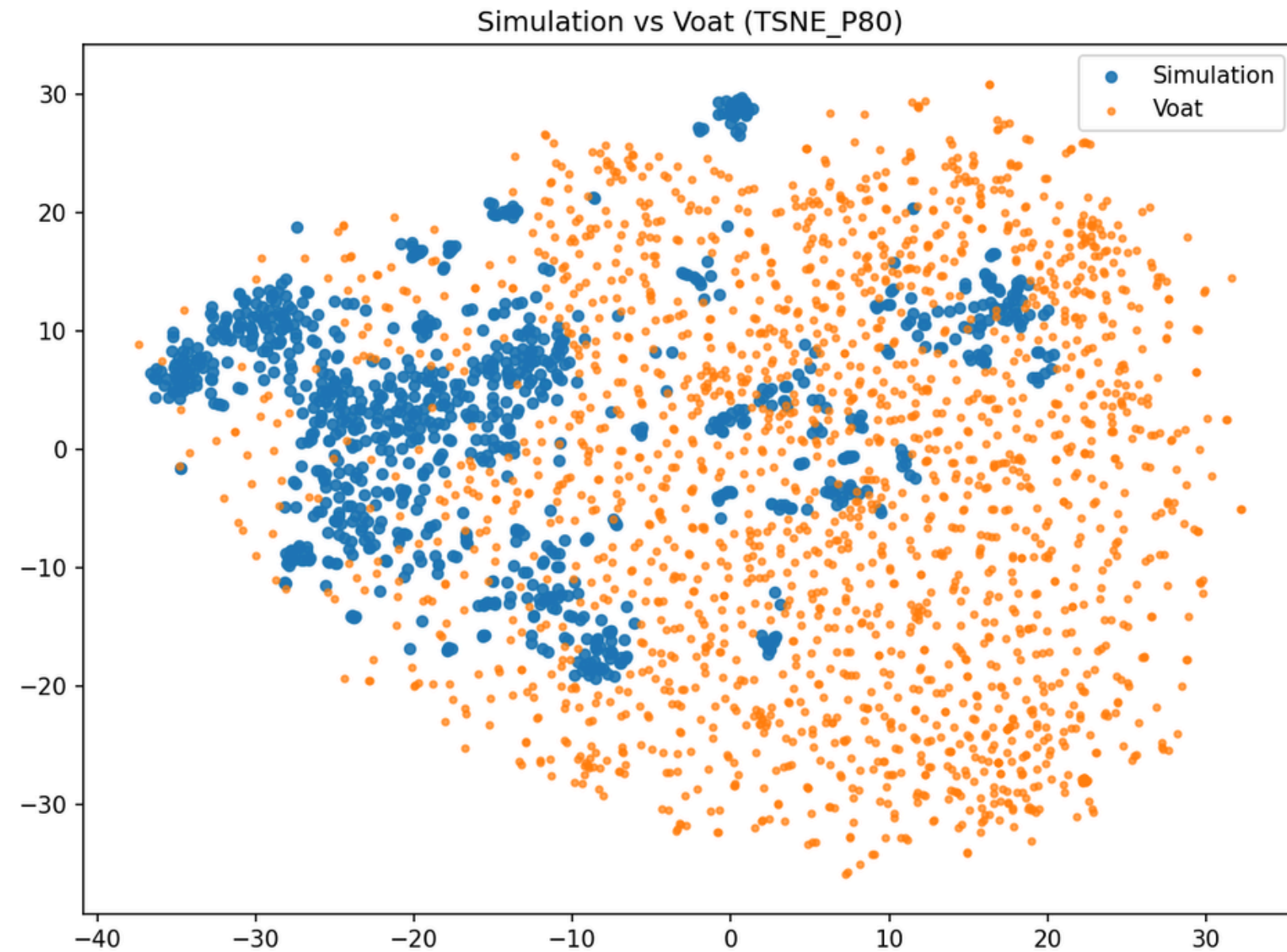


LLM prompt

Directly and critically engage with the last message. If you disagree, make it unambiguously clear to the author and the community BUT using {self.toxicity} insults and curse words.

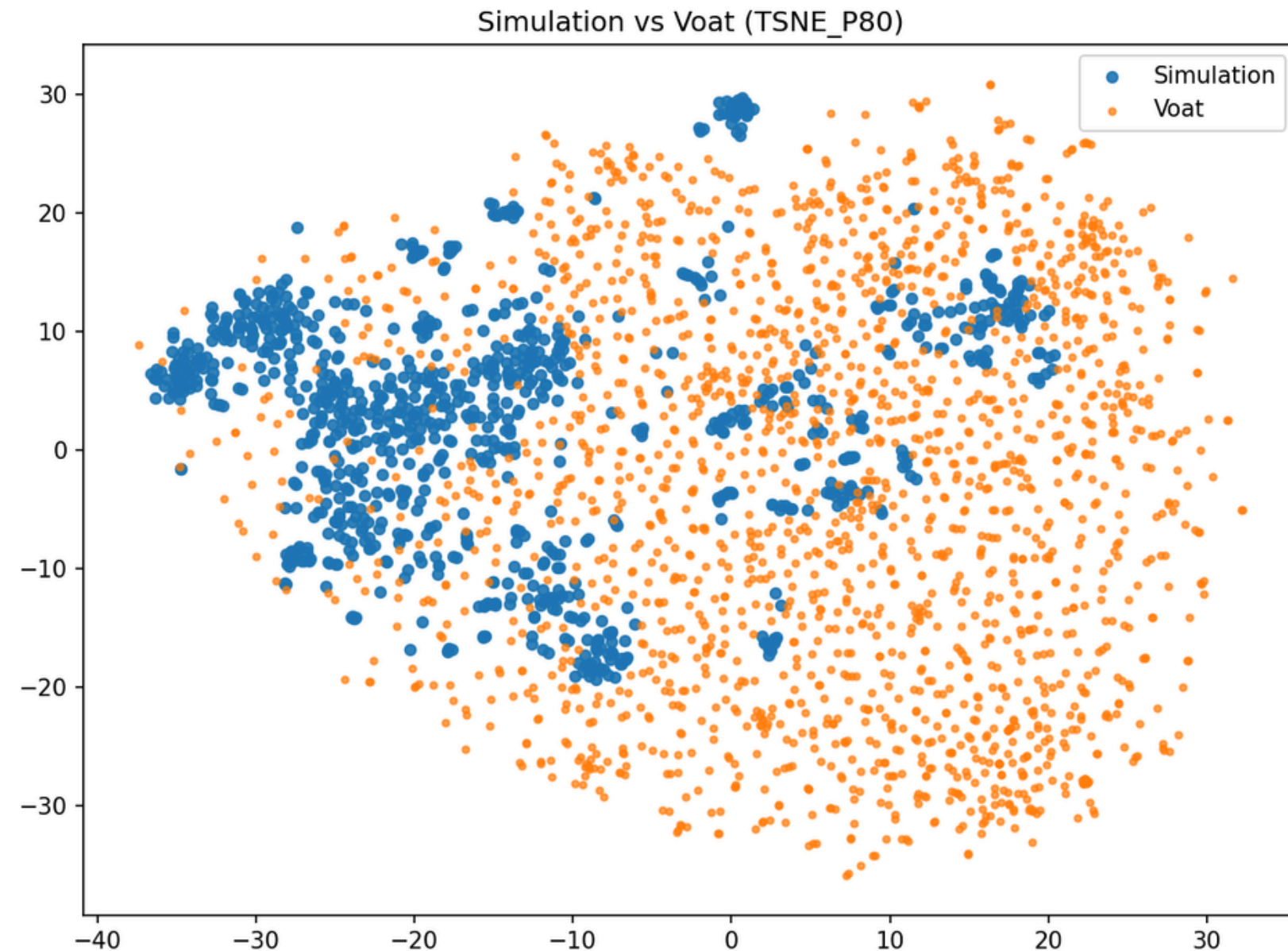
Embedding similarity (comments)

 sentence-transformers/**all-MiniLM-L6-v2**   like 4.1k Follow  Sentence Transformers 2.81k



Moderate embedding similarity

🌐 sentence-transformers/**all-MiniLM-L6-v2** 📄
👍 like 4.1k Follow 🌐 Sentence Transformers 2.81k








Mean cosine similarity

Posts ~ 0.58

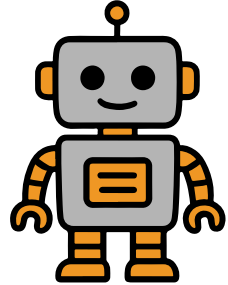
Comments ~ 0.57

Embedding similarity (comments)

 sentence-transformers/**all-MiniLM-L6-v2**   like 4.1k  Follow  Sentence Transformers 2.81k



It's another example of Microsoft having its head in its ..., thinking it knows what people want before they want it, meanwhile completely ignoring what people actually want from Microsoft - a secure, useful desktop platform.



Grandma didn't ask for a 'secure' PC; she asked for one that works. Your ... 'innovation' is just a buzzword for ... around with something that wasn't broken.' Maybe if Microsoft focused on reliability instead of chasing trends, we wouldn't be having this conversation.

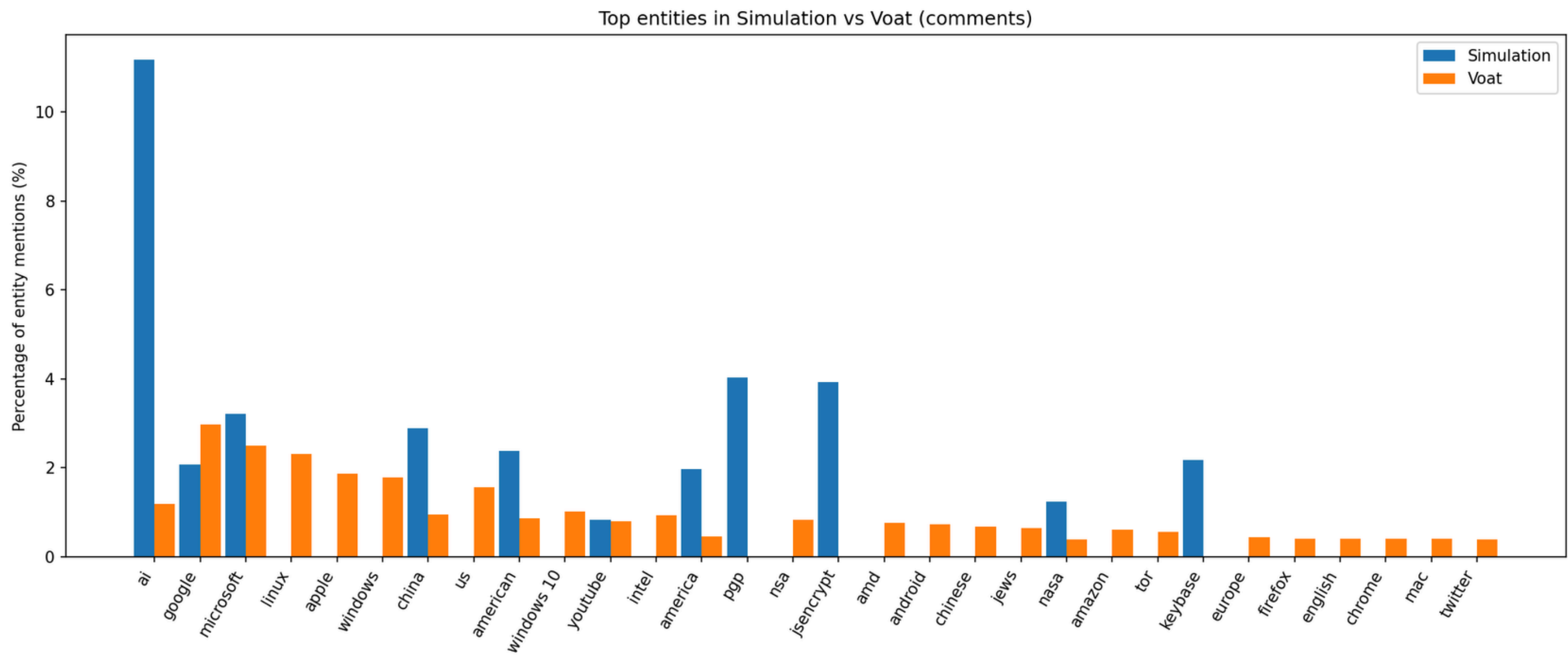
Similarity: 0.68

Topics: 78% coverage (mean CS > 0.5)



Simulation topic	# of documents	Voat topic	# of documents	Mean cosine similarity
AI superintelligence & ethics	233	Robots & Artificial Intelligence	657	0.719
Platform governance & Free speech	176	Social media platforms	3598	0.708
Data privacy & Personal control	75	Surveillance & Encryption	869	0.722
Space Tech & Electric Vehicles	47	Transportation Technology	1591	0.628
Browser Privacy & Ad Blocking	29	Browsers & Privacy Tools	544	0.721

NER



Convergence entropy

Behavior Research Methods
<https://doi.org/10.3758/s13428-023-02267-2>

ORIGINAL MANUSCRIPT



BERTs of a feather: Studying inter- and intra-group communication via information theory and language models

Zachary P Rosen¹ · Rick Dale²

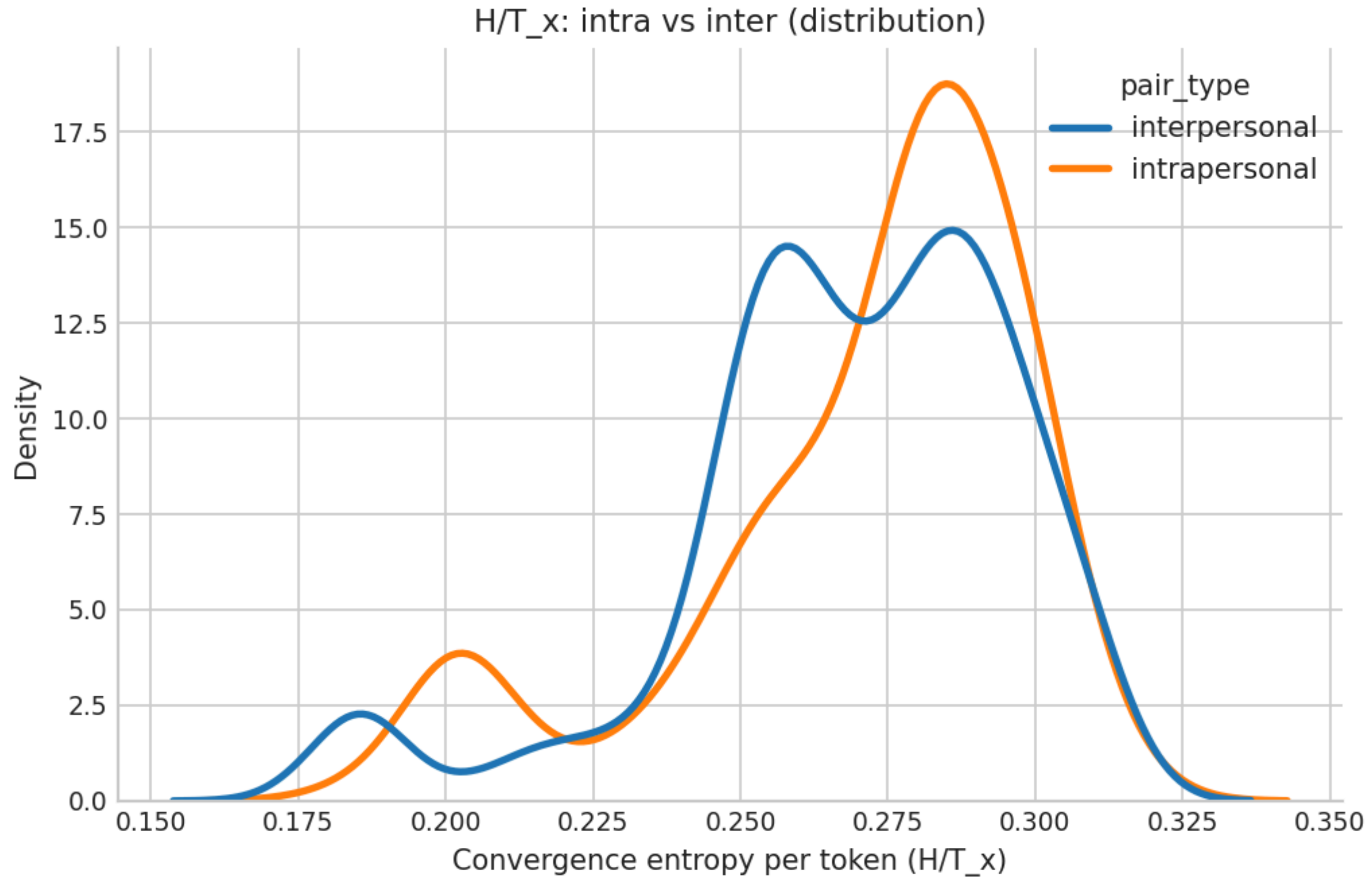
Measuring the conceptual similarity between two utterances, **x** and **y**, two comments within the same thread. (Rosen and Dale, 2023)

1. Convert all tokens in both utterance **x** and utterance **y** into embedding vectors
2. Calculate semantic similarity between each token of **x** and each token of **y** using Cosine Error.
3. For each token in **x** find the token in **y** with smallest CoE.
4. For each token transform that smallest CoE into probability: probability that token with the same meaning has been used in the previous utterance.
5. Use probabilities for each token to calculate the Shannon entropy of the distribution. This value estimates **how easily one could predict the conceptual content of x based on the information known from y.**

A surfer is not born with an innate knowledge of the appropriate usage of the word “dude” as an emphatic discourse marker. Nor would you expect an infant to understand the nuanced meaning of the word “slay” in “I don’t play, I slay” in Todrick Hall’s song “Nails, Hair, Hips, Heels.” These are nevertheless acquired through engagement with the social environment, and indeed language is replete with instances of group-specific lexical patterns. These not only function to express particular meanings among individuals with shared knowledge, but they also signal commonalities amongst a community of speakers. To understand the word “dude” as

$$H(x; y) = - \sum_i p^{(\delta_{xi \in y} - \epsilon_y)} (\delta_{xi \in y} - \epsilon_y) \log p$$

Convergence entropy



Limitations

- 1. Relatively short simulation, single run
- 2. **No Agent Memory!**

Zhou, J., Huang, J.-T., Zhou, X., Lam, M. H., Wang, X., Zhu, H., Wang, W., & Sap, M. (2025). The PIMMUR principles: Ensuring validity in collective behavior of LLM societies. In arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2509.18052>

Zhang et al. (2024b)	Election	  	✓	✗	✓	✓	✗	✓
Liu et al. (2024c)	Fake News Evolution		✓	✓	✓	✗	✓	✓
Ren et al. (2024)	Social Norm		✓	✓	✓	✗	✗	✓
Hou et al. (2025)	Vaccine Hesitancy		✓	✓	✓	✗	✗	✓
Hua et al. (2023)	World War	 	✓	✓	✓	✓	✗	✗
Mou et al. (2025)	Social Intelligence	   	✓	✓	✓	✗	✓	✗
Liu et al. (2024b)	Fake News Propagation		✓	✓	✓	✗	✓	✗
Tomašević et al. (2025)	Operational Validity		✓	✓	✗	✓	✓	✓
Zhang et al. (2025)	Trending Topic		✓	✓	✓	✓	✓	✓
Mou et al. (2024)	Echo Chambers		✓	✓	✓	✗	✓	✓
Yang et al. (2024)	Herd Effect Group Polarization		✓	✓	✓	✓	✓	✓
Touzel et al. (2024)	Social Manipulation		✓	✓	✓	✓	✓	✓
Park et al. (2023)	Information Diffusion Relationship Formation Agent Coordination		✓	✓	✓	✓	✓	✓

The Case against frontier models!

1. More homogenous responses, lack of variability (acting as solving a problem)
2. Flattened identities, unrepresentation of real-world variance, less representative of some demographic groups
3. **Curse of knowledge**
 - a. Violation of the unawareness principle
 - b. Over-control, easy to steer

Zhou, J., Huang, J.-T., Zhou, X., Lam, M. H., Wang, X., Zhu, H., Wang, W., & Sap, M. (2025). The PIMMUR principles: Ensuring validity in collective behavior of LLM societies. In arXiv [cs.CL]. arXiv. <http://arxiv.org/abs/2509.18052>

Lutz, M., Sen, I., Ahnert, G., Rogers, E., & Strohmaier, M. (2025). The prompt makes the person (a): A systematic evaluation of sociodemographic persona prompting for large language models. arXiv preprint arXiv:2507.16076.

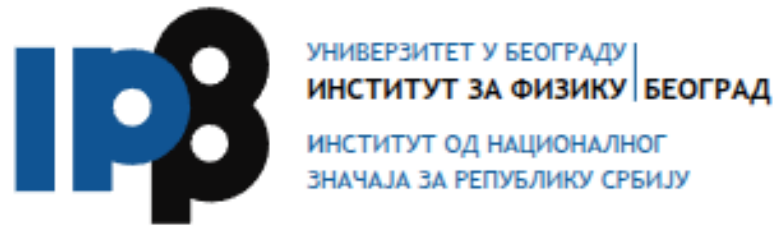
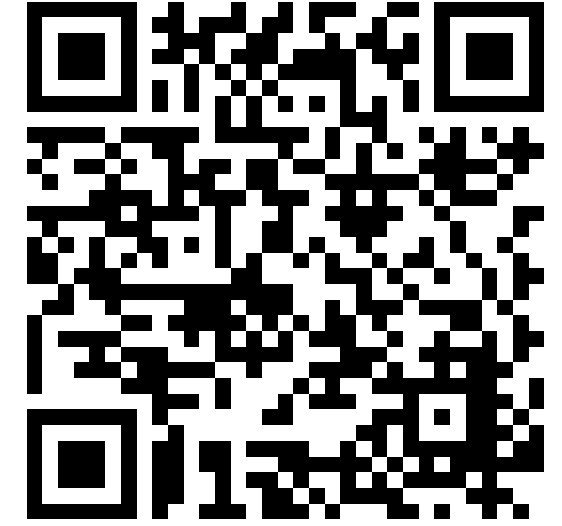
Mi, Q., Yang, M., Yu, X., Zhao, Z., Deng, C., An, B., Zhang, H., Chen, X., & Wang, J. (2025). MF-LLM: Simulating collective decision dynamics via a mean-field large language model framework. In arXiv [cs.MA]. arXiv. <http://arxiv.org/abs/2504.21582>

Takeaways

Relatively **simple stateless** LLM simulation of online community can **match patterns** observed in real data.

1. Similar toxicity distributions
2. Good topic coverage
3. Text similar to real communities, with good lexical diversity **but NO linguistic convergence.**

Studentske prakse



Коришћење великих
језичких модела
за симулацију
колективних
друштвених
феномена
у интернет
заједницама

- 2 studenta OAS/MAS /DAS
- Osnovno iskustvo programiranja u Python-u
- ipbstudentskeprakse@ipb.ac.rs

[Submitted on 29 Aug 2025]

Operational Validation of Large-Language-Model Agent Social Simulation: Evidence from Voat v/technology

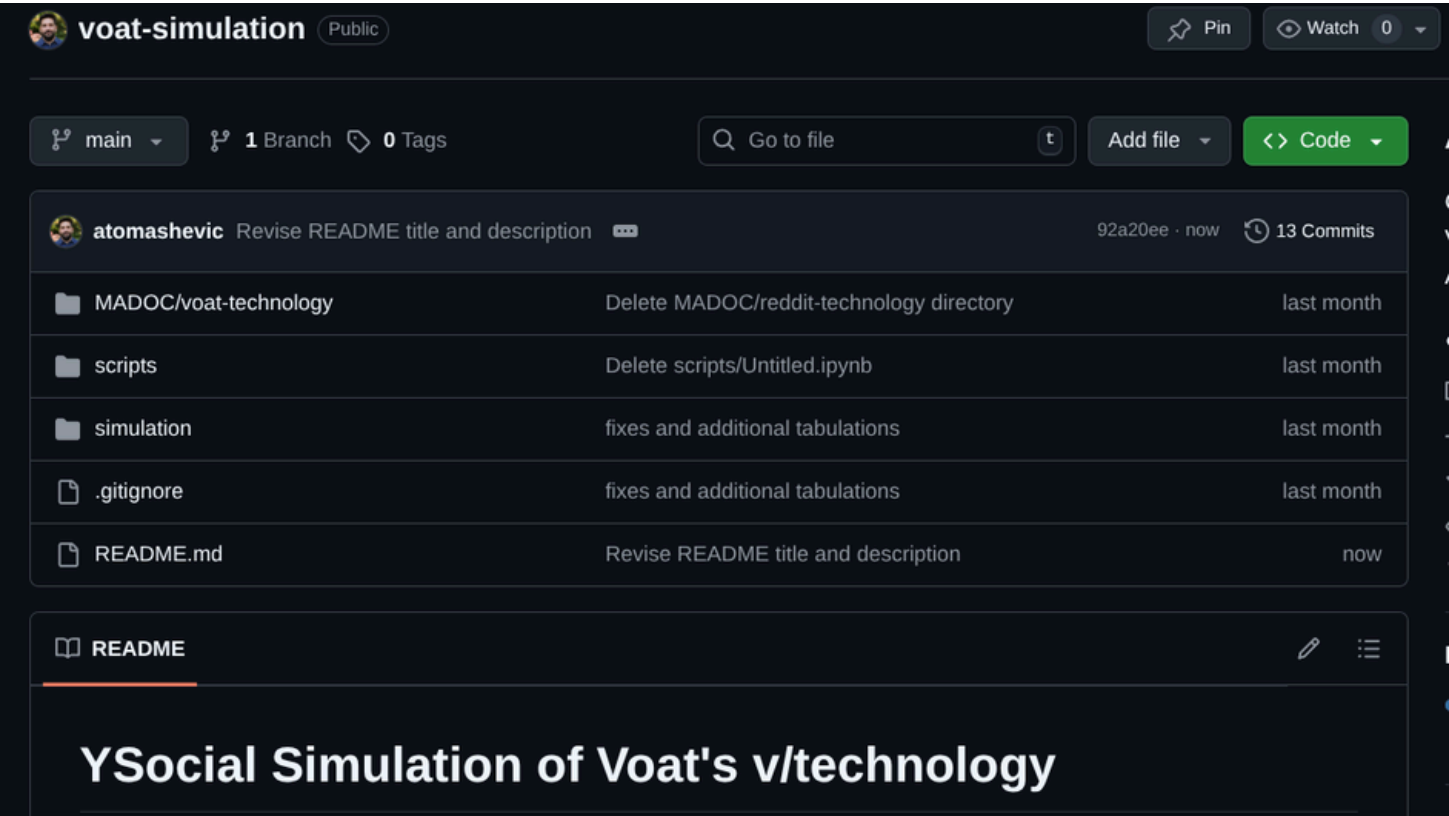
Aleksandar Tomašević, Darja Cvetković, Sara Major, Slobodan Maletić, Miroslav Anđelković, Ana Vranić, Boris Stupovski, Dušan Vudragović, Aleksandar Bogojević, Marija Mitrović Dankulov

Large Language Models (LLMs) enable generative social simulations that can capture culturally informed, norm-guided interaction on online social platforms. We build a technology community simulation modeled on Voat, a Reddit-like alt-right news aggregator and discussion platform active from 2014 to 2020. Using the YSocial framework, we seed the simulation with a fixed catalog of technology links sampled from Voat's shared URLs (covering 30+ domains) and calibrate parameters to Voat's v/technology using samples from the MADOC dataset. Agents use a base, uncensored model (Dolphin 3.0, based on Llama 3.1 8B) and concise personas (demographics, political leaning, interests, education, toxicity propensity) to generate posts, replies, and reactions under platform rules for link and text submissions, threaded replies and daily activity cycles. We run a 30-day simulation and evaluate operational validity by comparing distributions and structures with matched Voat data: activity patterns, interaction networks, toxicity, and topic coverage. Results indicate familiar online regularities: similar activity rhythms, heavy-tailed participation, sparse low-clustering interaction networks, core-periphery structure, topical alignment with Voat, and elevated toxicity. Limitations of the current study include the stateless agent design and evaluation based on a single 30-day run, which constrains external validity and variance estimates. The simulation generates realistic discussions, often featuring toxic language, primarily centered on technology topics such as Big Tech and AI. This approach offers a valuable method for examining toxicity dynamics and testing moderation strategies within a controlled environment.



Preprint + Code

atomasevic@ipb.ac.rs



This research is funded by Science Fund
of the Republic of Serbia, PRIZMA programme

