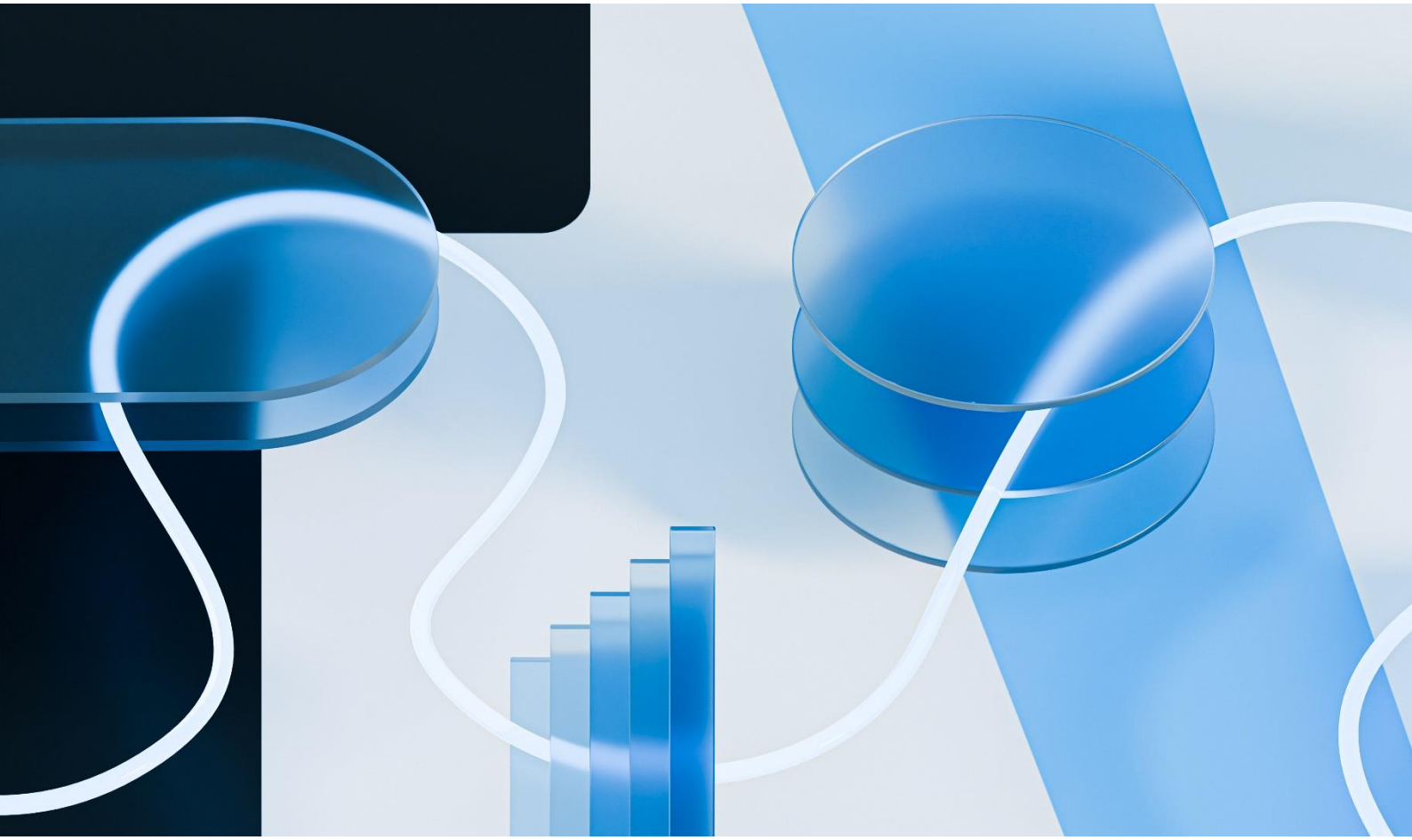
twon
twin of online social networks



TWin of Online Social Networks

Deliverable D4.4

TWON Computational report

Main Authors: Alenka Guček, Abdul Sittar

Funded by
the European Union

About TWON

TWON (project number 101095095) is a research project, fully funded by the European Union, under the Horizon Europe framework (HORIZON-CL2-2022-DEMOCRACY-01, topic 07). TWON started on 1 April 2023 and will run until 31 March 2026. The project is coordinated by the Universiteit van Amsterdam (the Netherlands) and implemented together with partners from Universität Trier (Germany), Institut Jozef Stefan (Slovenia), FZI Forschungszentrum Informatik (Germany), Karlsruher Institut für Technologie (Germany), Robert Koch Institute (Germany), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (Serbia), Slovenska Tiskovna Agencija (Slovenia), Dialogue Perspectives e.V (Germany).

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Funded by
the European Union

Project name	TWin of Online Social Networks
Project acronym	TWON
Project number	101095095
Deliverable number	D4.4
Deliverable name	Computational report
Due date	31.03.2026
Submission date	27.03.2026
Type	Report
Dissemination level	Public
Work package	WP4
Lead beneficiary	Institut Jozef Stefan, JSI
Contributing beneficiaries and associated partners	Universität Trier (UT), Karlsruher Institut für Technologie (KIT), Univerzitet u Begogradu - Institut za Filozofiju I Drustvenu (UoB)

Executive summary

This report presents a technical overview of the computational simulations developed within the TWin of Online Social Networks (TWON) project. The primary focus is on a data-driven multi-agent framework that leverages Large Language Models (LLMs) to simulate realistic social media ecosystems. This framework integrates three critical components: a Temporal Fusion Transformer (TFT) to predict realistic posting rhythms, persona-based agents to emulate diverse user behaviours, and Retrieval-Augmented Generation (RAG) to ensure conversational grounding.

The simulations were evaluated across three major topical domains—Technology, Climate Change, and COVID-19—comparing synthetic interactions against real-world Reddit data. Results indicate that while the system successfully reproduces structural network dynamics, such as bursty activity and information cascades, maintaining deep semantic coherence in long-term multi-turn discussions remains a significant challenge.

Furthermore, the report synthesises findings from various TWON implementations, including simple-user models for studying opinion dynamics, modular data-driven layers for testing LLM robustness, and persona-validated agents benchmarked against human participants. A key cross-cutting insight is the impact of ranking strategies on network evolution; for instance, engagement-based ranking was found to consistently increase influence inequality, while hybrid ranking can accelerate the formation of echo chambers. These results provide a validated, controlled environment for researchers to evaluate platform design choices and systemic risks without the ethical concerns of live-platform experimentation.

Contents

Tables

Figures

Abbreviations

LLM	Large Language Model
LoRA	Low-Rank Adaptation
LSS	Large Scale Simulation
OSN	Online Social Network
RAG	Retrieval-Augmented Generation
TFT	Temporal Fusion Transformer
TWON	Twin of an Online Social Network

Introduction

Deliverable D4.4, the TWON Computational Report, provides a technical overview of the large-scale simulations and computational experiments conducted within the TWin of Online Social Networks (TWON) project. This report is designed to be in sync with D2.2 (Report on Simulation Model); the two documents are strictly complementary. While D2.2 focuses on the theoretical foundations and the conceptual architecture of the models, D4.4 details their practical computational execution, parameter configurations, and the empirical results derived from large-scale runs.

The central principle of the TWON project is that the research question determines the model. Therefore, these simulations are not general-purpose tools but targeted instruments designed to enable counterfactual analysis, evaluate platform design choices, and assess systemic risks such as polarization and misinformation. Because different research objectives require different levels of abstraction, this report covers a spectrum of modelling granularities:

Abstract Formal Models: Such as the Simple-user TWON (KIT), which prioritizes computational tractability to isolate the specific effects of ranking algorithms on opinion dynamics.

Data-Driven LLM Systems: Such as the JSI, UT, and Belgrade TWONs, which utilize Large Language Models (LLMs) to simulate the linguistic and cognitive complexity of real human discourse.

This deliverable synthesizes the work of four consortium partners—KIT, JSI, UT, and Belgrade—structured around the core pillars of the User Model, Platform Model, and Technical Infrastructure. By documenting the computational performance and structural realism of these diverse implementations, D4.4 provides the evidence base for understanding how digital twins can reliably replicate the complex dynamics of online social ecosystems.

Simulation frameworks

The TWON project employs a modular, purpose-driven approach to simulation, adhering to the principle that the research question determines the model. This deliverable serves as a computational companion to D2.2 (Report on Simulation Model); this section details the practical technical frameworks used to execute large-scale experiments.

1. The Unified TWON Architecture

All simulations within the consortium are structured around three core pillars to ensure technical consistency across different research objectives:

User Model: Defines agent behavior, ranging from abstract update rules (e.g., Bounded Confidence Model) to generative AI personas that emulate human linguistic styles.

Platform Model: Represents the digital environment, including the social network topology (static or dynamic) and the ranking algorithms that curate the content visible to users.

Technical Infrastructure: The computational backend, typically orchestrated in Python, utilizing MongoDB for state persistence and High-Performance Computing (HPC) or cloud-based GPU clusters for execution.

2. Modeling Granularity and Objectives

The project balances a fundamental trade-off between computational tractability and behavioural realism. This has resulted in two distinct framework types:

Abstract Formal Models (e.g., Simple-user TWON): These models, such as the implementation by KIT, prioritize experimental control. By using simplified agents, researchers can perform large-scale parameter sweeps (e.g., 19,000 total simulation runs) to isolate how specific ranking algorithms drive polarization independently of user psychology.

Data-Driven Generative Models (e.g., JSI, UT, and Belgrade TWONs): These frameworks prioritize conversational realism. They utilize Large Language Models (LLMs) to simulate the cognitive richness of real discourse, grounding agent behaviour in empirical data from platforms like Reddit and X.

3. Shared Technical Standards for Generative TWONs

The generative frameworks share several advanced technical components designed to improve structural and linguistic realism:

Generative Backends: Primary models include LLaMA-2-7B and Llama-3.1-8B, often deployed as 8-bit quantized versions to optimize GPU memory.

Parameter-Efficient Fine-Tuning (PEFT): The use of LoRA (Low-Rank Adaptation) allows for domain-specific specialization (e.g., Technology or COVID-19) without the cost of full model training.

Temporal and Contextual Grounding: A Temporal Fusion Transformer (TFT) is frequently used as a learned scheduler to predict realistic posting rhythms. To prevent semantic drift, Retrieval-Augmented Generation (RAG)—supported by FAISS vector indexing—selects semantically relevant historical context for each agent interaction.

4. Cross-Framework Ranking Paradigms

A central objective across all frameworks is evaluating the impact of content exposure on network dynamics. The frameworks implement diverse ranking strategies to test for filter bubbles and influence inequality:

- Chronological Ranking: Acts as a baseline, prioritizing the most recent content.
- Engagement-Based Ranking: Prioritizes content with high interaction counts (likes/comments), often leading to rich-get-richer effects.
- Hybrid/Semantic Ranking: Combines engagement signals with social or ideological proximity (e.g., Semantic Similarity Ranker), which has been shown to accelerate modular segregation and echo chamber consolidation.

Data-driven Large Scale Simulations

This section presents the core computational experiments of the TWON project, focusing on the data-driven Large Language Model (LLM)-based multi-agent simulations. The objective is to evaluate how algorithmic curation and temporal dynamics jointly shape the evolution of online social networks, using a controlled digital twin environment grounded in real-world data.

The design and evaluation of these simulations follow the methodology described in Sittar et al (*Simulating Multi-Agent Social Media Ecosystem with LLMs: Modeling, Evaluation, and Insights on Filter Bubbles*, under review in Expert Systems Applications), which serves as the primary technical reference for this section.

See also Sittar el al (Constructing a Dataset to Support Agent-Based Modeling of Online Interactions: Users, Topics, and Interaction Networks, under review in IEEE Access) for the dataset construction that was used in LSS simulation.

For details on the platform that combines also real users see Sitar et al (*TWON: A Modular MERN-Stacked Based Platform for Social Intervention Studies,* under review in Social Network Analysis and Mining) and for experimetn with echo chambers a conference paper by Zust et al (Breaking Echo Chambers Through Diversity Injection: A Simulation Study of Interaction-Level Thresholds?, under review in International European conference on parallel and distributed computing).

1. Objective and Experimental Scope

The primary objective of these simulations is to evaluate the structural realism of an LLM-based ecosystem and the degree of algorithmic bias introduced by different exposure mechanisms. Architecture can be found in Figure 1. The experiments were conducted across three topical domains derived from real Reddit data: Technology (33 agents), Climate Change (14 agents), and COVID-19 (7 agents), see Figure 2, where the size of each bubble reflects the agent's level of activity, measured as a weighted sum of posts and comments.

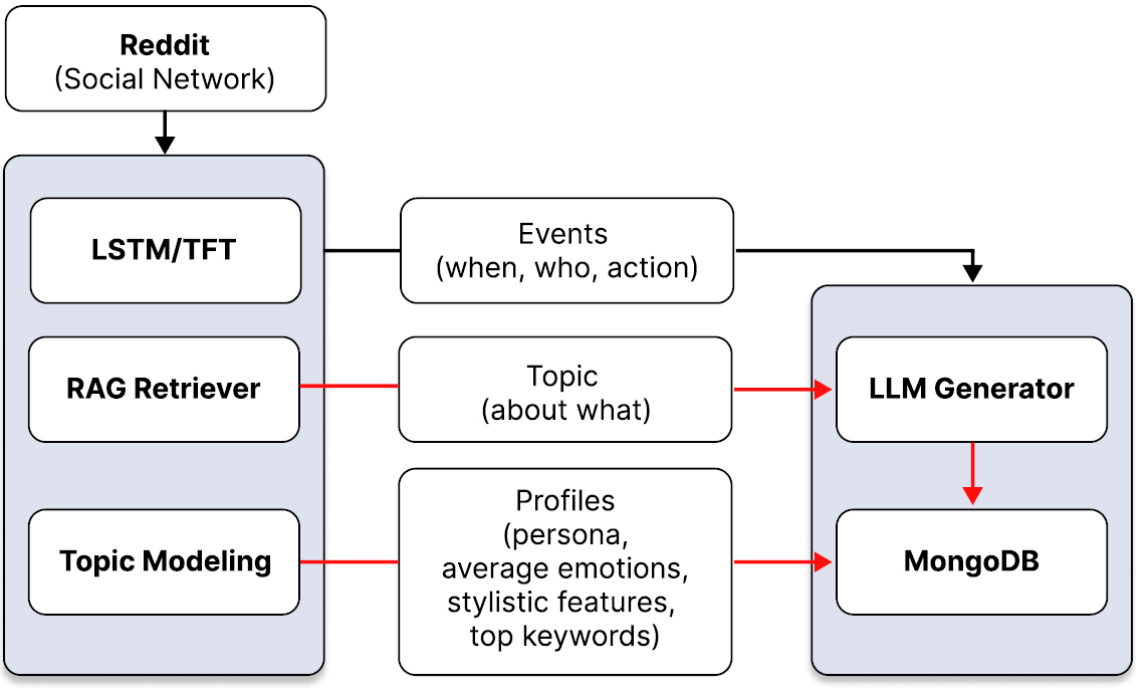


Figure 1: Overview of the LLM-based multi-agent social media

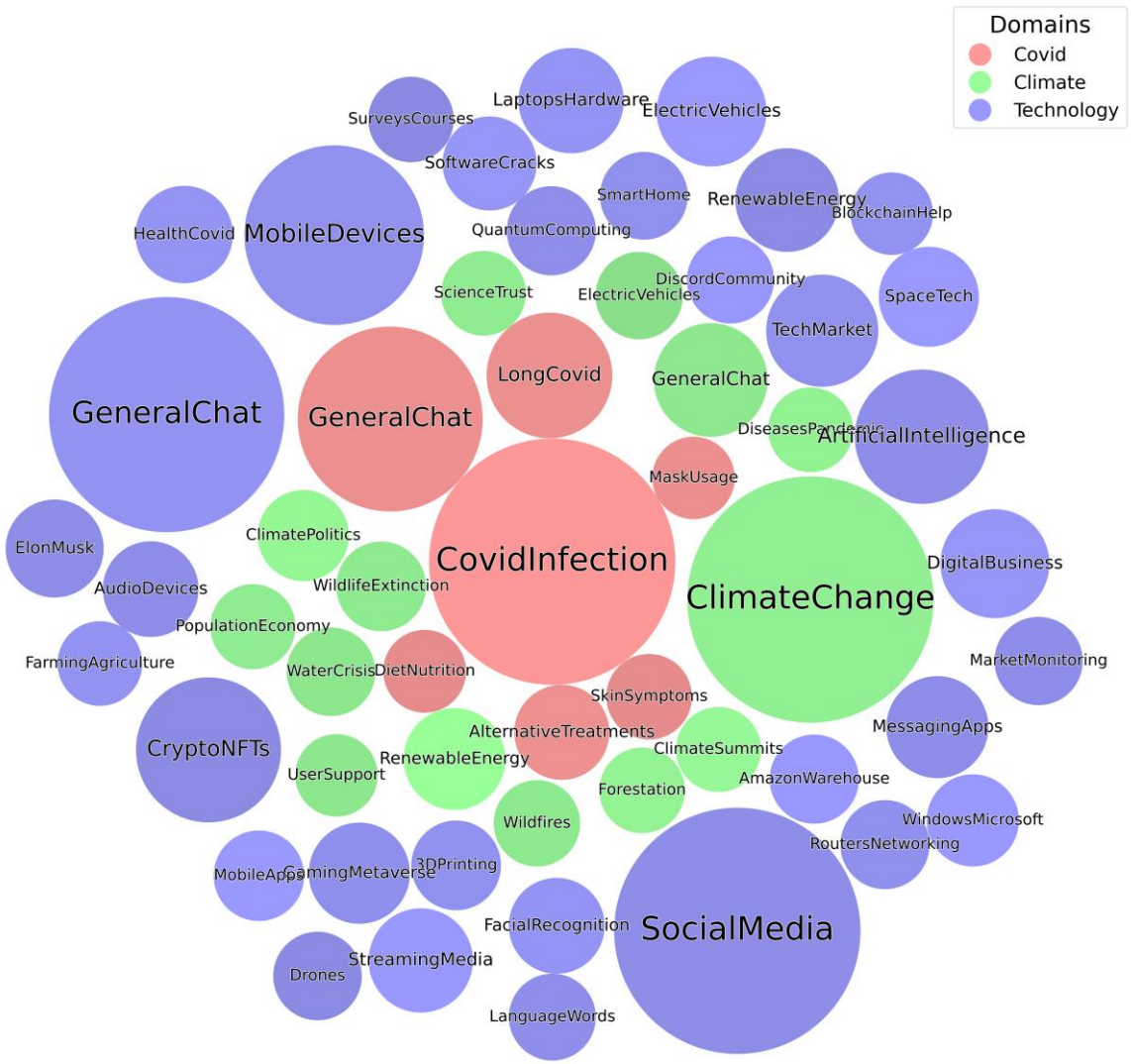


Figure 2: Visualization of all agents: 33 technology-focused, 14 climate-focused, and 7 COVID-related agents.

The simulation runs were executed at three distinct scales to observe the transition from baseline timing to feedback-driven network growth:

- Baseline: 100 events, reproducing real temporal sequences without ranking to isolate the effect of timing on realism.
- Mid-scale: 500 events (100 real initial events followed by 400 synthetic events), introducing ranking feedback to shape cascade growth.
- Large-scale: 900 events, incorporating dynamic follow-tie inference to allow the network topology to evolve alongside exposure mechanisms.

See Table 1 for the overview snapshot of the Data-driven LSS and Table 2 for Computational report.

Table 1: Experiments and Computational Details for the Data-driven LLM (JSI) TWON

Category	**Data-driven LLM TWON (JSI)**
Purpose	Bridge computational social science and generative AI using data-driven LLM agents to study filter bubbles and echo chambers.

Platform Model	Real Reddit discussion networks (Tech, Climate, COVID). Six ranking strategies: global and local variants of chronological, engagement-based, and hybrid exposure.
User Model	Learned scheduler (TFT) predicting next action (who, when, what). Content generated using LLaMA-2-7B with LoRA fine-tuning and retrieval-augmented generation (RAG).
Results	Ranking algorithms shape emergent structure; engagement-based ranking increases influence inequality; hybrid ranking strengthens clustering and potential echo chamber formation.
Technical Details	**Infrastructure:** Python, MongoDB, FAISS, 8-bit quantized LLaMA-2-7B with gradient accumulation. **Agents:** 7 to 33 aggregated community agents representing thousands of real users. **Interactions:** 100 (Baseline), 500 (Mid-scale), 900 (Large-scale) events per run. **Duration:** 47–87 min (Baseline), 134–379 min (Mid-scale), up to 855 min (Large-scale Tech).

Table 2: Computation Report for the Data-driven LLM TWON (JSI)

Category	Details
Objective	Evaluate realism and structural bias in LLM-based ecosystems using data-driven generative agents deployed on real discussion network topologies.
Platform Configuration	Real Reddit interaction graphs (Tech, Climate, COVID domains). Six ranking strategies: global and local variants of chronological, engagement-based, and hybrid exposure mechanisms.
Agent Architecture	Aggregated community agents (7–33 per domain). Temporal Fusion Transformer (TFT) scheduler predicting interaction target and timing. Content generation via 8-bit quantized LLaMA-2-7B with LoRA fine-tuning and retrieval-augmented generation (RAG).
Experimental Manipulations	Topical domain (3), ranking strategy (6), simulation scale (Baseline: 100 events; Mid: 500 events; Large: 900 events).
Evaluation Metrics	In-degree and out-degree centrality, interaction type distribution, sentiment evolution, topic drift, structural clustering, and influence inequality.
Number of Runs	105 manually evaluated interaction chains; additional large-scale synthetic runs across all ranking configurations.
Simulation Scale	Events per run: 100 (Baseline), 500 (Mid-scale), 900 (Large-scale).
Runtime Characteristics	Baseline: 47–87 min; Mid-scale: 134–379 min; Large-scale: up to 855 min (Tech domain).
Infrastructure	Python-based orchestration; MongoDB persistence; FAISS vector retrieval; 8-bit quantization with gradient accumulation for memory-efficient inference.

2. Experimental Setup: Ranking and Exposure

The simulations utilized six variations of ranking strategies as the primary independent variables to test "what-if" scenarios regarding content exposure:

- Chronological (Global/Local): Acts as a control, prioritizing the most recent content to preserve interaction diversity.
- Engagement-based (Global/Local): Prioritizes high-activity threads, testing for the "rich-get-richer" effect and increased influence inequality.
- Hybrid (Global/Local): Combines engagement signals with social proximity, evaluated for its potential to accelerate modular segregation and echo chamber formation.

See Table 3 for the ranking effects across domains where global and local chronological, engagement, and hybrid rankers are compared against baseline simulations using qualitative and quantitative evaluation metrics.

Table 3: Effects of ranking strategy and simulation scale on conversational structure and interaction patterns across domains.

Baseline simulations										
Settings		Qualitative evaluation				Quantitative evaluation				
Domain	Ranker	Cascading structure	Sentiment evolution	Conversational dynamics	Topic drifting	Interaction type distribution (synthetic vs real)	Agent population size	Influence concentration	Agent identity overlap	Simulation time (minutes)
COVID	TFT	no	no	no	no	2/98 (comments, posts) 59/41 (comments, posts)	3/7 (synthetic/real)	CovidTherapies/GeneralChat	2	87
Climate	TFT	no	no	no	no	0/100 68/32	6/9	ClimateChange/GeneralChat	4	49
Technology	TFT	yes	yes	yes	yes	74/26 52/48	15/18	DigitalBusiness/GeneralChat	9	47
Mid-scale simulations										
COVID	Chronological (G)	no	no	yes	yes	12/488	4/7	CovidTherapies/GeneralChat	3	379
COVID	Chronological (L)	no	no	yes	yes	13/487	4/7	CovidTherapies/GeneralChat	3	147
COVID	Engagement (G)	yes	no	yes	yes	12/488	4/7	CovidTherapies/GeneralChat	3	296
COVID	Engagement (L)	no	no	no	no	7/493	4/7	CovidTherapies/GeneralChat	3	190
COVID	Hybrid (G)	no	no	yes	yes	8/492	4/7	CovidTherapies/GeneralChat	3	326
COVID	Hybrid (L)	no	no	yes	yes	7/493	4/7	CovidTherapies/GeneralChat	3	171
Climate	Chronological (G)	no	no	no	no	1/499	11/9	ClimateChange/GeneralChat	6	275
Climate	Chronological (L)	no	no	no	no	2/498	10/9	ClimateChange/GeneralChat	6	549
Climate	Engagement (G)	no	no	no	no	1/499	11/9	ClimateChange/GeneralChat	7	285
Climate	Engagement (L)	no	no	no	no	3/497	9/9	ClimateChange/GeneralChat	5	191
Climate	Hybrid (G)	no	no	no	no	0/500	9/9	ClimateChange/GeneralChat	5	544
Climate	Hybrid (L)	no	no	no	no	1/499	12/9	ClimateChange/GeneralChat	7	198
Technology	Chronological (G)	yes	yes	yes	yes	111/389	32/18	StreamingMedia/GeneralChat	17	274
Technology	Chronological (L)	yes	yes	yes	yes	215/784	32/18	StreamingMedia/GeneralChat	18	379
Technology	Engagement (G)	yes	yes	yes	yes	109/391	28/18	StreamingMedia/GeneralChat	16	281
Technology	Engagement (L)	yes	yes	yes	yes	81/419	32/18	StreamingMedia/GeneralChat	17	189
Technology	Hybrid (G)	yes	yes	yes	yes	86/414	28/18	StreamingMedia/GeneralChat	16	134
Technology	Hybrid (L)	yes	yes	yes	yes	102/398	32/18	StreamingMedia/GeneralChat	18	344
Large-scale simulations										
Technology	Chronological (G)	yes	yes	yes	yes	176/824	32/18	StreamingMedia/GeneralChat	18	855
Technology	Chronological (L)	yes	yes	yes	yes	215/784	32/18	StreamingMedia/GeneralChat	18	379
Technology	Engagement (G)	yes	yes	yes	yes	203/797	33/18	StreamingMedia/GeneralChat	18	302
Technology	Engagement (L)	yes	yes	yes	yes	186/814	32/18	StreamingMedia/GeneralChat	18	828
Technology	Hybrid (G)	no	no	yes	yes	12/792	33/18	StreamingMedia/GeneralChat	18	758
Technology	Hybrid (L)	no	no	yes	yes	16/884	29/18	StreamingMedia/GeneralChat	16	363

3. Quantitative Evaluation: Structural Realism

Both qualitative and quantitative assessments were performed, see Figure 3 for details. The quantitative analysis focused on the system's ability to replicate real-world network metrics and temporal patterns.

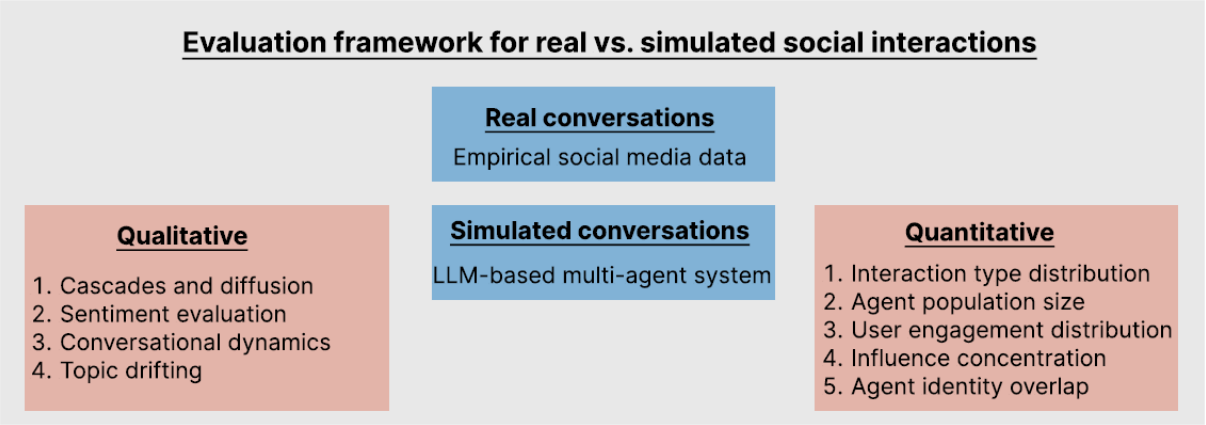


Figure 3: Evaluation framework for real vs. simulated social interactions.

Scheduler Performance: A critical finding was that the Temporal Fusion Transformer (TFT) significantly outperformed traditional LSTM models in predicting next-event attributes (actor, action, target, and timing), providing the necessary backbone for realistic activity bursts.

Domain Sensitivity: Structural realism was found to be highly domain-dependent. The Technology domain demonstrated the strongest realism, successfully reproducing cascading structures and interaction distributions that closely matched empirical data.

Ranking Effects: Results confirmed that engagement-based ranking consistently increases influence gaps, while hybrid ranking strengthens within-group reinforcement, a pattern consistent with echo chamber consolidation.

4. Qualitative Assessment: Conversational Dynamics

To complement the structural metrics, a manual qualitative evaluation of 368 interaction chains was conducted in the Technology domain.

Qualitative evaluation confirms that the simulations reproduce common interaction patterns such as question–answer exchanges and information cascades. However, longer interaction chains reveal a limitation: semantic coherence degrades over time, with topic drift and repetition emerging in extended discussions. This highlights a gap between structural realism and conversational depth.

5. Outcomes and Key Insights of the Data-driven TWON

The large-scale simulations conducted by JSI yielded three primary insights regarding the evolution of online social ecosystems. First, temporal modelling is the foundational prerequisite for structural realism. The experiments demonstrated that when the scheduler—specifically the Temporal Fusion Transformer (TFT)—accurately predicts the timing and targeting of events, the simulation successfully reproduces real-world patterns like bursty activity and information cascades. This is most evident in the Technology domain, where the TFT achieved significantly higher predictive calibration than traditional LSTM models.

Second, the simulations revealed that ranking mechanisms are the primary drivers of network evolution, dictating how influence is distributed across the platform. The outcomes of the "what-if" scenarios showed distinct trajectories for each strategy:

Chronological Ranking: Acted as a stabilizing force, preserving interaction diversity and limiting extreme influence concentration.

Engagement-Based Ranking: Systematically amplified popular threads, thereby increasing influence inequality and reinforcing "rich-get-richer" dynamics.

Hybrid Ranking: By prioritizing social proximity, this strategy significantly increased within-cluster interactions. In large-scale runs (900 events), this led to modular segregation, where information became trapped in tightly connected clusters, indicating the early stages of echo chamber consolidation.

Finally, the results highlight a "realism gap": while the framework achieves high structural plausibility (network metrics and timing), conversational realism remains a challenge. Qualitative analysis of 368 interaction chains showed that semantic coherence degrades in extended threads (50+ comments), which often devolve into paraphrasing or topic drift.

Summary of other Large-Scale Simulations

While the primary focus of this report is the data-driven generative framework, the TWON project encompasses several other modelling approaches tailored to specific research questions. These simulations range from abstract formal models designed for high-volume parameter exploration to persona-validated systems benchmarked against human behavioural data. This section synthesizes the computational efforts of the KIT, UT, and Belgrade partners.

1. Simple-user TWON (KIT): Isolating Algorithmic Effects

The Simple-user TWON prioritises computational tractability and experimental control over linguistic realism. By replacing LLMs with simple update rules based on the Bounded Confidence Model (BCM), this framework enables massive-scale experimentation, including 19,000 total simulation runs (see Table 4).

- Key Findings: The simulation demonstrated that ranking algorithms can override individual psychology; specifically, the "Closest" (filter bubble) ranker sustains high levels of polarization even when users are open-minded.
- Computational Scope: Each run involves 1,000 agents across 2,000 timesteps, executed in parallel on high-performance computing (HPC) clusters.

Table 4: Experiments conducted with the Simple-user (KIT) TWON

Category	Details
Manipulations	Agents' confidence bound ϵ (10 values, 0–0.4); ranking algorithm (Random, Engagement × 5, Closest, Narrative × 3 targets, Diverse Engagement, User Success × 5).
Outcome variables	Final polarisation (4× variance), network homophily, filter bubble strength.
Scale	1,000 agents per run; 2,000 timesteps per run; 19,000 total simulation runs.
Computational Details	**Infrastructure:** Python, NumPy, igraph; executed on HoreKa HPC cluster (KIT) via SLURM. **Compute time:** ≈ 10 hours per job; Total CPU hours ≈ 1,600.

2. Modular Data-Driven TWON (UT): Robustness and OOD Prediction

The UT TWON utilizes a modular, dual-layer architecture to evaluate the operational robustness of LLMs, specifically Llama-3.1-8B, in predicting unseen behavioural trajectories. See Table 5 for overview of experiments, Table 6 for computational details and Table 7 for computational report.

- Key Findings: Initializing agents with authentic user histories significantly improves empirical realism and ensures stable performance even in out-of-distribution (OOD) scenarios (e.g., transitioning from general X data to COVID-19 contexts). However, allowing too much "dynamic cognition" update can introduce behavioural drift.
- Computational Scope: Simulations involve 1,000 to 4,000 empirically seeded agents executed over 288 discrete timesteps, utilizing elastic cloud-based GPU clusters.

Table 5: Overview of Experiments Conducted with the Modular Data-Driven TWON (UT)

Experiment	Details
Experiment 1: Robustness of LLMs as backends in ABMs in Social Simulations	**Manipulations:** 3 Agents Setups: A) Naive Prompting (No sophisticated persona, short memory, static cognition) B) Data-Driven (Inferred persona profile, inferred activation patterns, longer memory, static cognition) C) Over-Aligned (Inferred persona profile, inferred activation patterns, longer memory, dynamic cognition). Contexts: In-distribution vs. Out-of-distribution (COVID scenario). **Outcome variables:** Similarity of synthetic feed with test feed on linguistic (Nela, Spacy) and semantic features (TweetEval) **Number of runs:** 1 **Number of agents:** 2000 (Consistent across runs) **Number of steps:** 288 discrete simulation timesteps **Duration of experiment:** ≈ 7-8 hours per batch of 4 concurrent runs (using 5 GPUs)
Experiment 2: Impact of ranking mechanisms on discourse quality	**Manipulations:** Ranking Algorithm (ChronologicalRanker vs. Semantic-SimilarityRanker). **Outcome variables:** Sentiment, civility, interactivity and rationality **Number of runs:** 1 **Number of agents:** 2000 (Consistent across runs) **Number of steps:** 288 discrete simulation timesteps **Duration of experiment:** ≈ 7–8 hours per batch of 4 concurrent runs (using 5 GPUs)
Experiment 3: Institutional intervention and asymmetric information	**Manipulations:** Injection of an authoritative health provider account (e.g., RKI) vs. Baseline. Context: Feed initiated with COVID-related posts **Outcome variables:** Sentiment, civility, interactivity and rationality **Number of runs:** 1 **Number of agents:** 2000 (Consistent across runs)

Number of steps: 288 discrete simulation timesteps

Duration of experiment: ≈ 7–8 hours per batch of 4 concurrent runs (using 5 GPUs)

Table 6: Experiments and Computational Details for the Modular Data-Driven TWON (UT)

Category	Modular Data-Driven TWON (UT)
Purpose	Combine experimental flexibility with empirical realism. Tests LLM operational robustness and out-of-distribution prediction by replicating real-world X users in a controlled, data-driven simulation environment.
Platform Model	Configurable NetworkX topologies (Complete, Barabási–Albert). Content feed initialized with authentic data. Three modular ranking mechanisms: Random, Chronological, and Semantic Similarity (via embeddings).
User Model	Empirically grounded stochastic activation (based on authentic posting frequencies). Content generation and evaluation mediated by LLMs using sliding memory windows (parameter instantiated with 15 actions) and authentic interaction histories.
Results	Empirically initializing LLM agents with authentic user histories maximizes both in-distribution fidelity and out-of-distribution generalization, whereas giving too much degrees of freedom to the LLMs (i.e. dynamic cognition updating) introduces behavioural drift without improving predictive accuracy.
Technical Details	**Infrastructure:** Python (abstract base classes), RunPod dynamic vLLM backend, 3–5 NVIDIA RTX 4090 GPUs. **Runs:** 1 per distinct configuration **Agents:** Default configuration of 1,000 - 4,000 empirically seeded LLM agents. **Interactions:** 288 discrete simulation timesteps per run (1 timestep equals ≈ 10 minutes) **Duration:** Variable, constrained by dynamic vLLM scaling and asynchronous API response latency. 4 simultaneous simulations with 2000 users and 4 GPUs take ≈ 7-8 hours.

Table 7: Computation Report for the Modular Data-Driven TWON (UT)

Category	Details
Objective	Evaluate LLM robustness and behavioural prediction capabilities. Test data-driven agent simulation derived directly from authentic X data.
Platform Configuration	Dynamically instantiated network structures (e.g., Complete, Barabási–Albert). Content exposure curated by interchangeable algorithmic rankers (Random, Chronological, Semantic Similarity).
Agent Architecture	Persona-driven WP3Agents wrapping a dynamic LLM API. Agents possess empirically derived activation probabilities, a parameterized context memory buffer, and prompt instructions seeded by/inferred from their real-world conversational histories.

Experimental Manipulations	Ranking algorithms (Chronological vs. Semantic), agent cognitive configurations (Naive vs. Data-Driven vs. Over-Aligned), In- and Out-of-Distribution contexts (Standard vs. COVID-19), and targeted injection of institutional personas (e.g., RKI).
Evaluation Metrics	Similarity of synthetic feed with test feed on linguistic (Nela, Spacy) and semantic features (TweetEval), Sentiment, Civility, Interactivity and Rationality.
Number of Runs	1 per distinct configuration
Simulation Scale	1000–4000 concurrent agents executing over 288 global timesteps. This results in ≈ 10,000 posts.
Runtime Characteristics	7-8 hours for 2000 Agents
Infrastructure	Python-based orchestration via YAML/CLI configuration. Generative backend powered by elastic RunPod vLLM deployments on 3–5 RTX 4090 GPUs. Output state persistence via JSON and YAML serialization.

3. Persona-Validated LLM TWON (Belgrade): Human-Agent Benchmarking

The Belgrade TWON focuses on validating whether LLM agents can serve as reliable proxies for human users while observing the effects of progressive algorithmic personalization. See Table 8 for overview of experiments, Table 9 for computational details and Table 10 for computational report.

- Key Findings: In a large-scale benchmark involving 1,511 human participants, LLM agents achieved 70.7% accuracy in predicting real-world social media reactions. Furthermore, as personalization increased across five stages, network modularity rose sharply from 0.22 to 0.68, indicating the structural formation of ideological clusters.
- Computational Scope: The system tested 27 different LLMs as generative backends, producing approximately 6.85 million reaction observations.

Table 8: Overview of Experiments: Persona-Validated LLM TWON (Belgrade)

Category	Experiment 1: Effect of Algorithmic Personalization (Simulation)	Experiment 2: Benchmarking Agent Accuracy (Validation)
Manipulations	**5 personalization stages** (0%, 25%, 50%, 75%, 100%); **3 primed topics** (Climate Change, Gaza, Ukraine); **5 post types** (Entertainment, News, Primed, Followed Agent, Unfollowed Agent).	**27 LLMs** × **3 persona types** (demographics, attitudes, full) × **56 posts** (entertainment vs. news; positive vs. negative valence).
Outcome Variables	**Engagement intensity** (*EIt*), **network modularity** (*Qt*), **affective polarization** (*APt*), follow dynamics, and sentiment variance.	**Hamming accuracy**, reaction-specific accuracy (like, dislike, comment, share, no reaction), ICC, and marginal/conditional *R*2.
Number of Runs	**10 Monte Carlo replicas** per configuration.	**1,511 human participants** × **81 agent versions** each.

Number of Agents	**100 agent prompts** per run.	**120,000+ unique agent-persona combinations**.
Number of User Interactions	~**6.85 million** individual reaction observations.	~**6.85 million** reaction predictions.
Duration of Experiment	**Minutes per configuration** for simulation; variable based on API latency for validation.	**Survey:** Nov 21–28, 2025; **API queries:** dependent on model latency.

Table 9: Experiments and Computational Details for the Persona-Validated LLM TWON (Belgrade)

Category	Persona-Validated LLM TWON (Belgrade)
Purpose	Combine simulation of algorithmic polarization with empirical validation of LLM agent behavioural fidelity. Tests whether persona-driven LLM agents can serve as reliable proxies for human social media users.
Platform Model	Dynamic follower network with 100 agents. Five-stage personalization trajectory (universal → fully personalized). Content types: Entertainment, News, Primed topics (Climate, Gaza, Ukraine), Agent posts.
User Model	Agents grounded in Big Five personality, political attitude, cognitive style from Serbian survey data (N=1,511). Reaction probability combines ideological (α=0.45) and emotional (α=0.55) alignment. Emotional contagion through sentiment-modulated content generation.
Results	Simulation: Modularity increased 0.22→0.68 across personalization stages. Engagement peaked at moderate personalization. Validation: 70.7% accuracy (Study 1), 67% accuracy binary task (Study 2). LLM choice strongest predictor. Positivity bias: +11pp accuracy for positive content.
Technical Details	**Infrastructure:** Python, NetworkX, 27 LLM APIs. **Runs:** 10 replicas per config (simulation); 1,511 participants (validation). **Agents:** 100 per simulation; 81 versions per participant. **Interactions:** ≈ 6.85M reaction observations. **Duration:** Minutes (simulation); variable by API (validation).

Table 10: Computation Report for the Persona-Validated LLM TWON (Belgrade)

Category	Details
Objective	Evaluate how algorithmic personalization drives polarization and validate LLM agent behavioural fidelity against human ground truth.
Platform Configuration	Dynamic directed graph with endogenous edge formation. Five-stage recommender with α increasing 0→2.0. Content exposure probability proportional to agent similarity.

Agent Architecture	Persona-driven agents with psychometric profiles (Big Five, political attitude, cognitive style). Reaction function: $P(Like)=\sigma(\beta\cdot\text{ideological alignment}+\beta\cdot\text{emotional alignment})$. 27 LLM backends tested.
Experimental Manipulations	Personalization level (5 stages), content type (5), primed topic (3), persona specificity (3), LLM model (27), post valence (positive/negative), persona-content alignment.
Evaluation Metrics	Network modularity Qt, affective polarization APt, engagement intensity EIt, Hamming accuracy, reaction-specific accuracy, ICC, marginal/conditional $R2$.
Number of Runs	Simulation: 10 replicas × 75+ configs. Validation: 1,511 × 81 × 56 observations.
Simulation Scale	100 agents per simulation. 120,000+ agent-persona combinations. ≈ 6.85M total observations.
Runtime Characteristics	Simulation: Minutes per configuration. Validation: Parallel API queries across 27 models.
Infrastructure	Python, NetworkX, multi-API integration (OpenAI, Google, xAI, Anthropic, Meta, Mistral, DeepSeek, Qwen, NVIDIA). Data via OSF.

Conclusions

The TWON Computational Report (D4.4) serves as a technical synthesis of the diverse simulation frameworks developed within the project. The simulations documented in this report demonstrate that the research question determines the model, with the consortium successfully balancing the trade-off between computational tractability and behavioural realism.

1. Technical Achievements and Structural Realism

The project has achieved a significant milestone in social simulation by successfully moving from abstract, rule-based agents to data-driven, language-capable entities. Key technical outcomes include:

- Temporal Grounding: The implementation of the Temporal Fusion Transformer (TFT) proved essential for structural realism. By accurately predicting when interactions occur and to whom they are targeted, the system successfully reproduced real-world bursty activity and information cascades.
- Persona Fidelity: Generative frameworks (JSI, UT, and Belgrade) demonstrated that LLMs can emulate nuanced social behaviours and stances. The Belgrade validation study confirmed that LLM-based agents can achieve over 70% accuracy in predicting human social media reactions, proving their utility as reliable proxies for human users.
- Modular Architecture: The development of modular "dual-layer" infrastructures allowed for counterfactual analysis and the isolation of specific causal mechanisms, such as the independent impact of ranking algorithms on opinion dynamics.

2. The Impact of Algorithmic Curation

Across all implementations—from simple-user models to complex generative ecosystems—the findings consistently highlight those ranking algorithms are primary drivers of network structure.

- Chronological Ranking remains the most effective at preserving interaction diversity.
- Engagement-Based Ranking systematically amplifies influence inequality and creates "rich-get-richer" dynamics.
- Hybrid and Semantic Ranking accelerate modular segregation, strengthen within-cluster density, and foster the structural conditions for echo chamber consolidation.

3. Challenges and Future Directions: The Realism Gap

While the project successfully reproduced network-level structural patterns, a significant "realism gap" remains between structural plausibility and conversational depth.

- Semantic Coherence: Qualitative analysis indicates that while simulated interactions appear realistic in short sequences, long-term multi-turn discussions often suffer from topic drift, paraphrasing, and a lack of cumulative argumentative development.
- Future Work: To address these limitations, future research should focus on enhancing agent memory systems, developing dynamic personas that evolve through interaction, and implementing content-aware ranking strategies that prioritize novelty and thematic variety.

In summary, the TWON project has provided a validated, controlled environment for exploring "what-if" scenarios regarding platform design and systemic risks. By documenting the computational performance and empirical validity of these models, this deliverable establishes a foundation for using digital twins to evaluate digital democracy and mitigate the risks of polarization and misinformation without the ethical concerns of live-platform experimentation.

twon

twin of online social networks

Contact us

Damian Trilling
Project Coordinator

+31 62 782 7904

d.c.trilling@uva.nl

University of Amsterdam
Postbus 15791
1001 NG Amsterdam

Funded by
the European Union