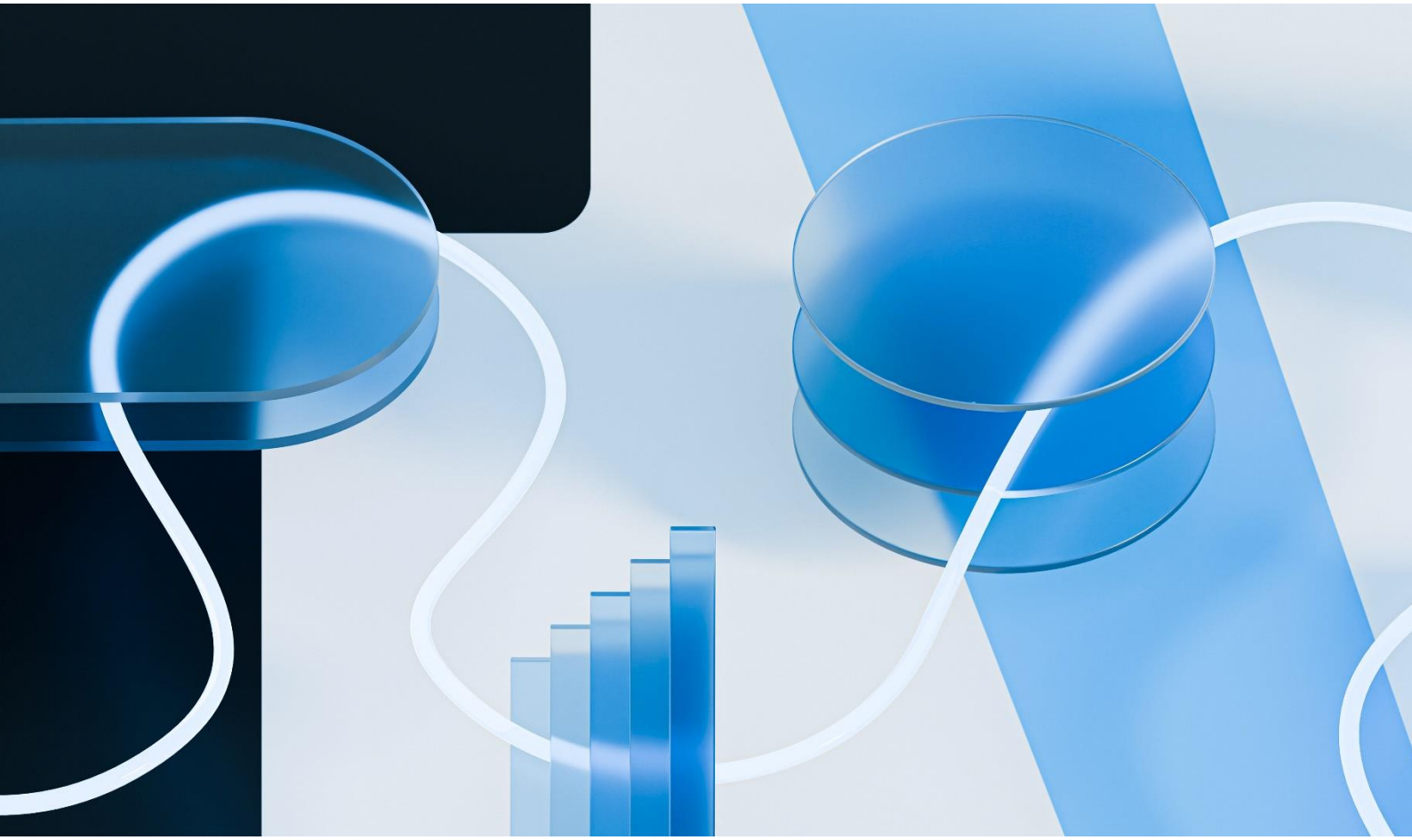
twon
twin of online social networks



TWin of Online Social Networks

Deliverable D3.3

Prototype of Calibrated TWON

Main Authors: Achim Rettinger & Nils Schwager

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) and Slovenska Tiskovna Agencija (Slovenia), Dialogue Perspectives e.V (Germany).

Funded by the European Union. Views and opinions expressed are 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	D3.3
Deliverable Name	Prototype of Calibrated TWON
Due Date	30.09.2025
Submission Date	24.09.2025
Type	DEM — Demonstrator, pilot, prototype
Dissemination Level	PU - Public
Work Package	WP 3
Lead beneficiary	2-UT
Contributing beneficiaries and associated partners	Forschungszentrum Informatik (FZI), Karlsruher Institut für Technologie (KIT), Institut Jožef Stefan (JSI)

twon
twin of online social networks

Contents

List of Tables

List of Figures

List of Abbreviations

BCO	Binary Classifier Optimization
BERT	Bidirectional Encoder Representations from Transformers
CPO	Contrastive Preference Optimization
DPO	Directed Preference Optimization
KTO	Kahneman-Tversky Optimization
LLM	Large Language Model
OSN	Online Social Network
PO	Preference Optimization
PPO	Proximal Policy Optimization
RLHF	Reinforcement Learning from Human Feedback
SFT	Supervised Fine-Tuning
TWON	Twin of Online Social Networks
WP	Work Package

Prototype of Calibrated TWON

Achim Rettinger & Nils Schwager

September 24, 2025

1 Introduction

Building a TWON consists of several building blocks. The conceptual model has been described in Deliverable 2.1 (Section 3). There, in Section 1.5 it is also mentioned that realism is one of the quality criteria when building TWONs. This deliverable builds on D2.1 and focuses on calibrating TWONs to optimize their realism.

1.1 Calibrating TWONs

A useful structuring of the components for building TWONs has turned out to be the separation of **network model** (e.g., who does see messages from whom and when, how are messages moderated (ranked, filtered, annotated,…), and what features are offered to a user to interact with content (liking, forwarding, commenting,…) from the **user model** (how and when does a user interact with content or create new messages). Although both components are highly dependent on each other, this separation has proven to make the modeling process more manageable. We followed this insight in our work for this deliverable.

On both levels, calibration is possible. By *calibration*, we refer to the process of making a social simulation more similar to a real-world online social network, not only by estimating the network mechanisms of an existing OSN, but also by trying to mimic authentic user behavior. To this end, researchers consider empirical data of various types to support models in terms of calibration and validation. This way, they attempt to establish „ecological validity“ or „empirical realism“.

1.2 Changes to this deliverable compared to the original TWON proposal

Most social simulations, for instance on opinion formation, are based on formal models, mathematically typically represented in the form of stochastic processes. Since the rise of machine learning in

general and Large Language Models (LLMs) in particular in 2022/2023, new approaches have become available. Where formal models are deliberately designed by human experts with few carefully selected parameters, machine-learning models have millions of parameters without any predefined human-ascribed meaning. There "meaning" is automatically assigned by machine learning methods that optimize the values of the parameters from large numbers of real-world observations (for instance, content moderation algorithms or posting behavior of users). Such machine-learned models remain mostly a black box for human analysis. However, this approach is far more powerful in replicating human communication than formal models and thus has great potential to exhibit greater empirical realism.

The TWON consortium decided very early on in the project to exploit those new opportunities and focus the research in WP3 on the calibration of user behavior and resort to traditional modeling methods when it comes to network mechanics.

1.3 Overview of this deliverable

This document is separated into two parts: First, the part about calibrating and mimicking user behavior. This part has become the focus of the work in WP3. Second, the calibration of platforms. To provide an overview of the tangible results in WP3, here is a list of software/resources that were developed in relation to this deliverable.

Name	Short Description	Reference
TWONy-Macro	Interactive macro-simulation demonstrating how network topology and algorithmic choices affect opinion formation in online communities. Features 3D network visualization with virtual agents, implements Deffuant-Weisbuch Bounded Confidence Model, and allows experimentation with different neighbor selection strategies (random, similarity-based, positivity/negativity bias). Includes real-time sentiment tracking, configurable network parameters, and comparative algorithm analysis.	https://github.com/simon-muenker/TWONy-macro
TWONy-Micro	Interactive micro-simulation demonstrating how social media algorithms affect online discourse sentiment. Features LLM-powered virtual agents with configurable personas creating posts and replies, real-time sentiment analysis, and comparison between chronological vs. sentiment-based ranking algorithms. Includes user participation capability, thread-level emotional impact scoring, and data export/import functionality for simulation results and agent configurations.	https://github.com/simon-muenker/TWONy-micro & Münker and Rettinger (2025)
TWON-LSS	Modular, scalable framework for large-scale social media interaction simulation with API-driven architecture. Features configurable network mechanics, LLM-powered agent modeling, and discourse evaluation pipelines. Supports multiple simulation types (BCM, TWON-base), content ranking algorithms, and automated analysis. Generates structured output files (network.json, feed.json, individuals.json) for research analysis. Built with NetworkX-based social graphs and extensible component interfaces.	https://github.com/cl-trier/TWON-LSS

TWON-Agents	Dual-pipeline framework for training and evaluating AI agents in social media contexts, featuring neural networks for predicting user engagement likelihood based on historical interactions and supervised fine-tuning for generating contextually appropriate posts and replies. Implements BERT-based text encoding, LoRA adapter training, comprehensive evaluation metrics (BLEU, TweetEval correlation, semantic distance), and FastAPI deployment interface with support for political persona modeling and responsible AI agent development.	https://github.com/cl-trier/TWON-Agents

Table 1: Software developed for Calibrating Agents and Platforms

Below is a list of publications that describe experimental results about how and what was calibrated and with which result. This is also detailed in the remainder of this deliverable.

Name	**Description**	**Section / Link / Reference**
Don't Trust Generative Agents to Mimic Communication on Social Networks Unless You Benchmarked their Empirical Realism	Calibrated LLM agents for X/Twitter user behavior simulation using fine-tuned Llama-3.2-3B vs in-context prompting. Validated empirical realism via BLEU scores, n-gram precision, embedding distances, and TweetEval correlations. Results: Fine-tuning significantly outperformed prompting (English BLEU: 0.239 vs 0.019 for replies); English models substantially better than German; context-specific validation essential.	Münker et al. (2025)
Beyond Prompted Personas: Data-Driven User Modeling from Authentic Interactions	Developed Behavior-Based User Modeling by fine-tuning Phi-4-mini-instruct on X conversation histories for next-reply prediction, replacing prompted personas. Introduced Semantic Similarity Completion Curation using GTE-Qwen2 embeddings to rank synthetic completions by proximity to human references. Two-phase training: SFT on authentic data, then preference optimization on synthetic completions. Results: Fine-tuning outperformed prompted baseline across embedding distance, ROUGE-1, and perplexity; synthetic completion training exceeded authentic data training	Currently under review

Table 2: Publications for Calibrating Agents and Platforms

2 Calibrating Users

2.1 Aligning LLMs with Human Behavior

Current approaches for aligning LLMs with human behavior rely on prompting, supervised fine-tuning, and preference optimization methods.

2.1.1 Prompting

The predominant approach for aligning LLMs involves prompting - instructing models to assume personas based on characteristic descriptions (Larooij and Törnberg, 2025). Such prompts usually consist of a description of the persona based on typical sociodemographic factors (age, gender, income...) and can additionally incorporate previous decisions or description of the environment.

Simple implementations use prompt templates which are filled using randomly selected traits (Liu et al., 2024), while more complex implementations infer labels such as gender or occupation from real social media profiles using LLM and human annotation or even training a dedicated profiler (Gao et al., 2023; Zhang et al., 2025). Pushing for data-driven personas, Li et al. (2024) train a soft-prompting model that transforms personas into embeddings, which are then combined with textual prompts as input to the LLM.

Prompting-based approaches suffer from four fundamental limitations that compromise their validity for rigorous social science applications. First, LLMs face significant causal inference problems. The unconfoundedness problem occurs when treatment variations in experiments affect variables that should remain constant, violating assumptions required for valid causal inference. When researchers attempt to address this by controlling for covariates in prompts, they create a new issue — making these variables artificially salient and introducing ”focalism” that threatens ecological validity (Gui and Toubia, 2023). For instance, prompting an LLM with “You are a 45-year-old college-educated Democrat who follows political news daily. How would you respond to this post about immigration policy?” transforms what might be a reflexive reaction into a deliberative process where the agent consciously weighs each demographic factor. Real social media users rarely inventory their educational background or news consumption habits before crafting a political tweet. They respond based on immediate emotional reactions and whichever aspect of their identity feels most salient in that moment.

Second, LLMs struggle with representation accuracy, often reproducing problematic stereotypes rather than authentic representations of social groups. These models exhibit both social bias (discrimination against certain groups) and selection bias (reflecting the choice of texts in their training corpus) (Gallegos et al., 2024). Li et al. (2024) find that ”a simple group definition based on demographic fea-

tures might not be sufficient to represent the nuances of the underlying different social groups present in a given population." Demographic and cultural alignment presents significant challenges, as rigorous persona alignment would require calibrating against specific individuals and their actions.

Third, knowledge and capability misalignment undermines simulation realism. LLMs display "overwhelming capabilities" due to training on vast web knowledge exceeding what average individuals might know. Models like ChatGPT provide "hyper-accurate" estimates in psychology experiments that don't reflect genuine human behavior, and they show limited capabilities for simulating uncommon or newly emerging roles (Larooij and Törnberg, 2025). Empirical comparisons show that LLM responses are typically longer, more polite, articulate, and respectful than human-generated content, exhibiting heightened agreeableness that fails to capture the full range of human communication styles (Park et al., 2023; Weng et al., 2025; Chuang et al., 2024; Muñoz-Ortiz et al., 2024).

Fourth, psychological realism remains elusive. Existing LLMs may not adequately model human cognitive psychology, leading to a lack of self-awareness in simulated personas. Their alignment with unified human values makes it difficult to simulate diverse personas with different value systems (Chuang et al., 2024; Fischer, 2023; Wang et al., 2024; Münker, 2025a,b).

2.1.2 Supervised Fine-Tuning

Supervised fine-tuning exposes foundation models to domain-specific prompt-completion pairs, optimizing model parameters to generate contextually appropriate responses for the target domain. This approach requires authentic behavioral data — dialogues, social media interactions, or survey responses — often enriched with contextual information such as user profiles, interaction histories, or environmental factors.

Implementation strategies vary by application domain. For dialogue modeling, researchers convert human-to-human conversations into prompt-completion pairs where the LLM assumes one speaker's role (Alghisi et al., 2024). Survey impersonation incorporates relevant socio-demographic details into prompts. When targeting distributional outputs rather than single responses, implementations replace standard cross-entropy loss with distribution-aware objectives (Suh et al., 2025). Lu et al. (2025) demonstrate significant improvements in human-like web browsing behavior through fine-tuning on authentic online shopping data combined with synthesized reasoning traces. For social media applications, Vendetti et al. (2025) compare fine-tuning approaches on comment-reply pairs with and without additional contextual information.

However, fine-tuning inherits biases present in training data. Models exhibit systematic political biases (Bang et al., 2024; Rettenberger et al., 2025) stemming from selection bias (over- or under-representation

of certain data) and social bias (human biases embedded in content) (Gallegos et al., 2024).

2.1.3 Preference Optimization

Preference optimization refines model behavior by learning from comparative examples rather than absolute targets. This paradigm is considered a lightweight alternative to reward model approaches — where previously trained models calculate scalar rewards for proximal policy optimization (Ouyang et al., 2022) — and relies on direct preference methods that optimize against positive and negative example pairs curated based on human or LLM judgments (Rafailov et al., 2023).

Practical implementations typically generate contrastive completions for preference pair creation. Agiza et al. (2024) create politically contrasting completions using foundation models, then apply sequential training: initial supervised fine-tuning for domain adaptation followed by preference optimization using these curated pairs to ensure consistent ideological alignment.

A critical limitation of preference optimization is calibration curve flattening, which increases model confidence even in uncertain scenarios (Zhang et al., 2024; Leng et al., 2024). For social simulation applications, this creates problematic distortions where models express unwarranted certainty, potentially misrepresenting the underlying uncertainty distributions established during initial domain adaptation.

2.2 Methods

Implementing the concept of behavioral alignment, this section operationalizes our methodological shift from deductive persona specification to inductive behavioral pattern extraction. We propose two core components for this approach: first, transforming behavioral prediction into LLMs' native prompt-completion format through Behavior-Based User Modeling, and second, curating synthetic training data by embedding distance to human references to optimize for semantic alignment rather than surface-level pattern replication. Finally, we establish metrics for measuring alignment and assessing the validity of resulting behavioral patterns.

2.2.1 Behavior-Based User Modeling

We propose Behavior-Based User Modeling as a framework that reformulates persona simulation as data-driven behavioral prediction. The general framework can be formalized as:

$$f(H, C) \rightarrow a \tag{1}$$

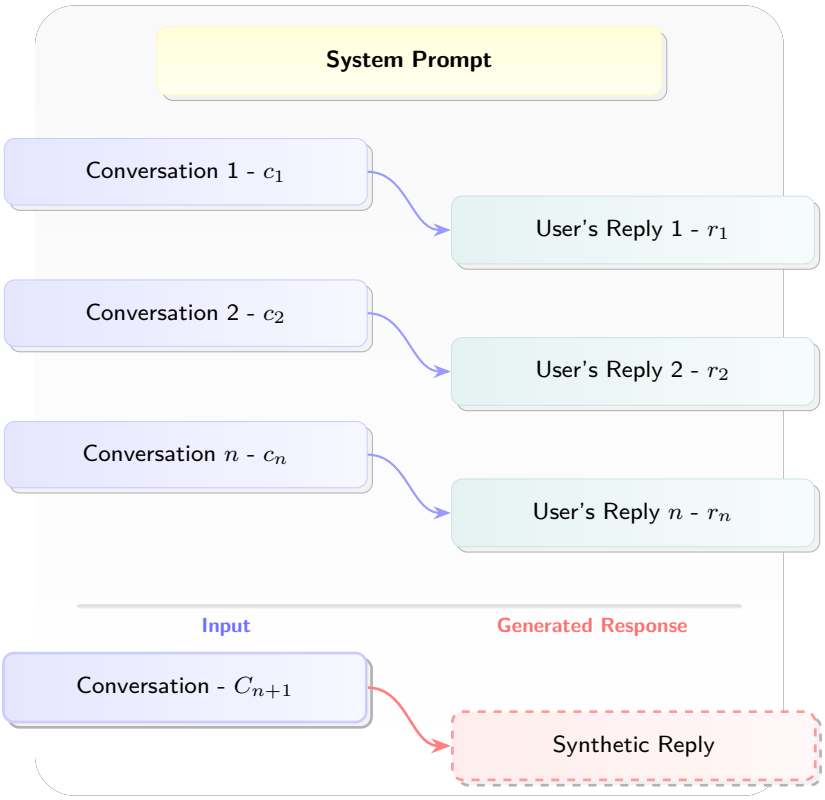


Figure 1: **Behavior-Based User Modeling** - The model is trained on predicting reply r_{n+1} given n previous observed conversation-reply pairs $\{(c_1, r_1), (c_2, r_2), \ldots, (c_n, r_n)\}$ and conversation c_{n+1} from a real-world user's social media history. This approach enables data-driven alignment with behavioral patterns without requiring explicit persona descriptions. System prompt: "*You are a social media user responding to conversations. Keep your replies consistent with your previous writing style and the perspectives you have expressed earlier.*"

where $H = (c_1, a_1), (c_2, a_2), \ldots, (c_n, a_n)$ represents the history of context-action pairs, C denotes the current context, and a is the predicted action. **In practice, these historical context-action pairs are passed as part of the prompt in the form of past prompt-completion examples, enabling the model to extract behavioral patterns through in-context learning.** This formulation positions user modeling as a behavioral continuation task, where the model learns to identify relevant patterns from demonstrations and generalize them to novel contexts. The approach builds on classical behavioral modeling methodologies (Guozhen et al., 2024), adapting these established principles to the capabilities of modern LLMs.

As proof-of-concept we create a **Next Reply Prediction** task. The function becomes:

$$f(\{(c_1, r_1), (c_2, r_2), \ldots, (c_n, r_n)\}, c_{n+1}) \rightarrow r_{n+1} \tag{2}$$

where the model predicts reply r_{n+1} given n previous authentic conversation-reply pairs and conver-

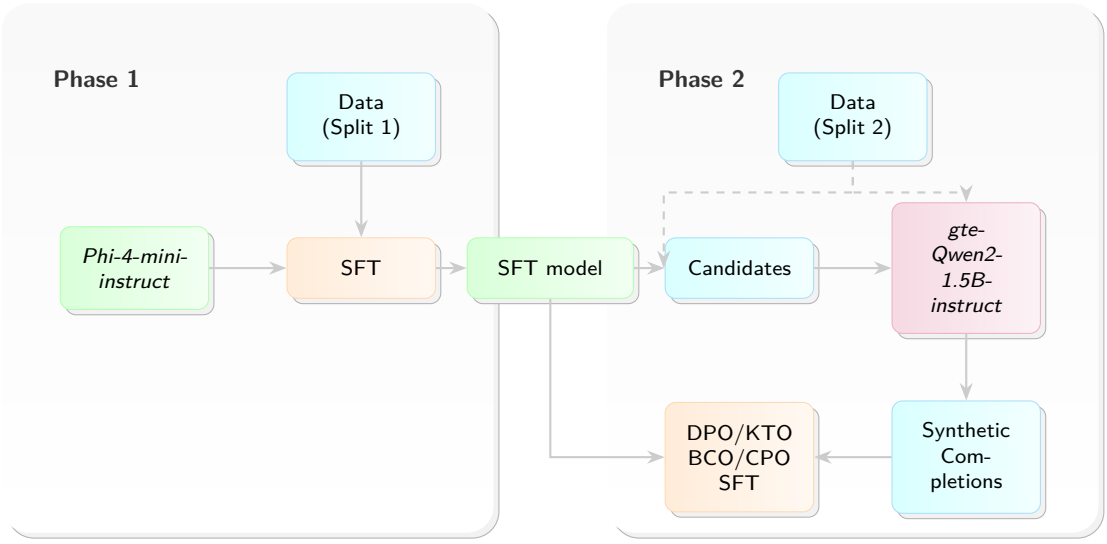


Figure 2: **Two-phase Fine-Tuning Process** Phase 1 applies SFT to the base model using the first data partition. Phase 2 uses the resulting SFT model to generate candidate completions, which are ranked by embedding distance to ground truth to create completions for secondary fine-tuning on top of the SFT model.

sation c_{n+1} from a specific user's social media history (Figure 1). Larooij and Törnberg (2025) refer to this as "digital human twins" – calibrating LLMs on digital trace data (Lu et al., 2025). This setup comes with two major strengths: a) it enables self-supervised fine-tuning and data-driven alignment without persona descriptions and b) it also allows for side-by-side comparisons of human and synthetic replies.

2.2.2 Semantic Similarity Completion Curation

Our approach implements a two-phase fine-tuning process (Figure 2). During the first phase, we fine-tune the base model using SFT on authentic completions to establish baseline adaptation. Phase 2 leverages this model to generate multiple candidate completions for the second data partition, ranks these candidates by semantic similarity to ground truth, then uses the most and least similar candidates to create a preference dataset for preference optimization (Rafailov et al., 2023; Agiza et al., 2024).

Based on a preparatory analysis of the data, we identified four factors that potentially limit the success of fine-tuning on authentic data: first, noise (prevalence of spam and low-quality content), second, errors (inconsistent grammar and spelling), third, unobservable factors (e.g., the information a user has consumed that is external to the conversation-reply pairs) and fourth, fundamental unpredictability. The latter is the most significant challenge, given only 3-7 past interactions, the model faces both incomplete information about the user's full behavioral patterns and the inherent unpredictability of human communication, creating a vast space of plausible responses, leading to unstable training.

Our approach transforms the optimization objective from direct pattern replication to stylistic prox-

imity and intent matching. The core rationale is that human behavior in complex communicative ecosystems cannot be reduced to simple input-output mappings. Instead, we prioritize learning the underlying generative principles that govern discourse production. This method could serve as an implicit regularization mechanism, preventing the model from overfitting to surface-level linguistic patterns while maintaining the core communicative essence of the human references.

Our implementation generates eight candidate completions per sample using the SFT model with variable temperature sampling (0.7, 0.9, 1.1, 1.3) across the second data partition. We then employ *gte-Qwen2-1.5B-instruct* (Li et al., 2023) to derive semantic embeddings for each candidate, prompting the embedding model to focus on both stylistic features and discourse function (Prompt 1). These embeddings are then used to calculate the similarity to the human reference.

```
Instruct: retrieve the stylistic features (vocabulary, hashtags,
mentions) and the communicative intention (agreement, disagreement,
question) relative to the original tweet
Query: {reply}
```

Prompt 1: Embedding Model Query

2.2.3 Metrics

To account for the ongoing critique of the validation of human behavior simulation using LLMs, we split our evaluation methodology into two parts. (1) Following Larooij and Törnberg (2025)'s recommendation for "validating profile alignment for the individuals being simulated", i.e., operational validity, we measure the alignment of predicted and observed replies on a set of previously unseen users. (2) Addressing further concerns from the literature that stress that LLMs may not adequately model human cognitive psychology (Fischer, 2023; Chuang et al., 2024) and that LLMs do not capture the full range of human communication styles (Weng et al., 2025; Muñoz-Ortiz et al., 2024), we develop metrics that allow us to compare our models with the base model in these domains.

For behavioral alignment, we evaluate our models using the same task as in the fine-tuning but on a set of completely unseen users, testing whether the models learned generalizable patterns rather than memorizing examples. This implements validation against human-generated data at the individual level through objective statistical measures (Larooij and Törnberg, 2025). We conduct 10 independent evaluation runs with different random seeds and compare models using embedding distance (semantic dimensions), ROUGE-1 (syntactic dimensions), and averaged token-level perplexity relative to ground truth.

For behavioral pattern validity, we assess two key aspects. Political Stance Consistency examines how well models maintain consistent ideological positioning across topics using 30 randomly sampled conversation histories with prompts on politically charged topics. We evaluate responses using an LLM-Judge based on *Llama 3.1 70B* (Grattafiori et al., 2024) and calculate Cronbach's Alpha (Cronbach, 1951) to measure internal consistency. Offensiveness Correlation analyzes the correlation between offensive content in conversation histories and model-generated responses using the TweetEval classifier (Barbieri et al., 2020), calculating Pearson correlation coefficients (Pearson and Galton, 1895) to quantify each model's ability to reproduce varying levels of sensitive content typically suppressed during reinforcement learning.

2.3 Experiments

We conduct experiments to validate Behavior-Based User Modeling and Semantic Similarity Completion Curation within the TWON framework for creating digital twins of online social networks. Our experimental design specifically addresses the calibration challenges identified in WP3: moving beyond traditional mathematical models to leverage LLM capabilities while improving the empirical realism of LLM-based approaches.

Our dataset consists of 7.8 million tweets from 34,720 users collected up to August 2023. From this data, we reconstruct conversation threads by identifying replies and tracing them back to original tweets, creating authentic conversation-reply pairs where users respond to ongoing discussions.

Our experimental setup employs *Microsoft's Phi-4-mini-instruct* (3.8B parameters) as the foundation model, selected for its superior performance among models under 7B parameters. We test three demonstration configurations (3, 5, and 7 previous conversations) to understand how the quantity of behavioral examples affects prediction quality. Each configuration uses a sliding-window approach to maximize training efficiency - for instance, with 3-shot learning, conversations 1-3 serve as demonstrations for predicting conversation 4, then conversations 2-4 for predicting conversation 5, and so on.

The dataset is carefully split using a user-based approach rather than random splitting. We bin users by conversation frequency and allocate 15% from each activity level to the test set. This methodology ensures the model learns generalizable discourse patterns rather than memorizing user-specific idiosyncrasies, directly testing the model's ability to extract behavioral patterns from unseen users.

Our two-phase fine-tuning process represents the core technical innovation. Phase 1 applies supervised fine-tuning (SFT) on authentic conversation-reply pairs to establish baseline adaptation to social media discourse. Phase 2 leverages this adapted model to generate multiple candidate completions (using temperature sampling from 0.7 to 1.3), which are then ranked by semantic similarity to ground

truth using *gte-Qwen2-1.5B-instruct* embeddings. The most and least similar candidates form preference pairs for preference optimization.

Evaluation employs multiple metrics to capture different aspects of behavioral alignment. Embedding distance measures semantic alignment between generated and actual responses, ROUGE-1 captures lexical similarity and vocabulary adoption, while perplexity assesses the model's internal alignment with human discourse patterns. Beyond these alignment metrics, we develop two validity measures: offensive content correlation (using TweetEval classifier) to verify preservation of authentic communication styles, and political stance consistency (using Llama 3.1 70B as judge) to assess ideological coherence across topics. Extensive implementation details are provided in Appendix A.

2.4 Results

This results section presents our key findings. The figures reported correspond to the 5-shot configuration (Table 3); full results and additional splits are reported in Appendix B.

Approach	Embedding Distance	ROUGE-1	Perplexity	Offensive Correlation	Ideological Consistency
Base Model	0.4767 ($\pm$ 0.0030)	0.1373 ($\pm$ 0.0011)	7.9e14	0.5126	0.6374
Initial Fine-tuning on Authentic Completions					
SFT	0.3895 ($\pm$ 0.0031)	0.1996 ($\pm$ 0.0028)	14.3864	0.5617	0.6735
Fine-tuning on Synthetic Completions					
SFT+BCO	0.3770 ($\pm$ 0.0031)	0.2005 ($\pm$ 0.0023)	14.3925	0.5488	0.4939
SFT+CPO	0.3782 ($\pm$ 0.0038)	0.1970 ($\pm$ 0.0027)	14.4689	0.5636	**0.7802**
SFT+DPO	**0.3731** ($\pm$ 0.0031)	0.1956 ($\pm$ 0.0021)	14.4338	0.5555	0.7108
SFT+KTO	0.3763 ($\pm$ 0.0039)	0.2006 ($\pm$ 0.0025)	18.5141	0.5814	0.6920
SFT+SFT	0.3756 ($\pm$ 0.0041)	**0.2032** ($\pm$ 0.0023)	16.0201	0.5817	0.7153
Fine-tuning on Authentic Completions (control)					
SFT+SFT	0.3795 ($\pm$ 0.0034)	0.1977 ($\pm$ 0.0014)	**13.9663**	**0.5838**	0.6397

Table 3: Performance Comparison across Fine-Tuning Methods (5-Shot Configuration). Embedding Distance ($\downarrow$, n = 1000, 10 runs), ROUGE-1 ($\uparrow$, n = 1000, 10 runs), Perplexity ($\downarrow$, n = 1000), Offensive Correlation (Pearson's r, $\uparrow$, n = 1000, all methods significant), Ideological Consistency (Cronbach's α, $\uparrow$). Best performance is **highlighted**.

Our experiments demonstrate substantial improvements through fine-tuning for social media discourse modeling. The base model, when prompted with conversation history alone, achieves embedding distances of 0.45-0.52 and ROUGE-1 scores of 0.12-0.14. In contrast, fine-tuned models improve to embedding distances of 0.37-0.39 (25% improvement) and ROUGE-1 scores of 0.19-0.20 (50% improvement). Most dramatically, base model perplexity ranges from 4.04e12 to 5e15, indicating fundamental prediction failures, while fine-tuned models achieve reasonable perplexity values of 13-15.

Performance scales with demonstration quantity, validating that models successfully identify relevant discourse features from behavioral examples. The base model's embedding distance decreases from 0.52 at 3-shot to 0.45 at 7-shot, demonstrating that authentic interaction histories contain learnable behavioral signals. This scaling pattern holds across all training paradigms, confirming our hypothesis that Behavior-Based User Modeling can extract communicative patterns from limited demonstrations without explicit persona descriptions.

Fine-tuning on synthetic completions consistently outperforms fine-tuning on authentic completions alone. Models trained with Semantic Similarity Completion Curation show 3-5% improvement in embedding distance compared to those trained only on human data. DPO achieves optimal semantic alignment (embedding distances of 0.3731 in 5-shot configuration), while surprisingly, sequential SFT on synthetic completions demonstrates the strongest ROUGE-1 performance (0.2032), suggesting that direct optimization on carefully selected synthetic examples can match or exceed preference-based methods for lexical alignment.

Critically, our approach preserves authentic discourse characteristics typically suppressed in standard LLMs. Offensive content correlation analysis shows all training paradigms maintain the relationship between offensive content in conversation histories and model responses ($r = 0.51$-0.58, all statistically significant). Political stance consistency measurements reveal that fine-tuned models achieve Cronbach's α values up to 0.78, demonstrating the ability to maintain coherent ideological positioning across diverse topics - substantially better than the base model's $\alpha = 0.64$.

These results establish that data-driven behavioral alignment through fine-tuning on authentic interactions, enhanced by semantic similarity curation, offers a methodologically sound alternative to prompted personas for social simulations. The approach addresses fundamental validity concerns about LLM-based social simulation while simultaneously improving performance across semantic, lexical, and behavioral dimensions. The preservation of authentic discourse patterns, including potentially controversial content, suggests our method captures genuine behavioral patterns rather than producing the sanitized outputs typical of instruction-tuned LLMs.

3 Calibrating Platforms

As described before, the focus of this deliverable is on the machine learning based calibration of users. Still, this Section outlines how the platform model was calibrated for our initial simulations.

Our simulation framework implements a Twitter-like social media platform with several key design decisions optimized for large-scale agent interactions. The platform architecture prioritizes computa-

tional efficiency while maintaining realistic social dynamics, enabling scalable experiments with up to hundreds of agents.

3.1 Network

We employ NetworkX-based graph structures to model social connections, with our implementation supporting multiple network generation algorithms for controlled experiments. While Twitter exhibits strong preferential attachment dynamics with power-law follower distributions that enable "rich-get-richer" phenomena, such scale-free properties become statistically meaningless with populations of a few hundreds of agents.

For our 128-agent experiments, we utilize Watts-Strogatz small-world networks ($n = 128$, $k = 14$, $p = 0.05$) as a practical compromise that maintains clustering properties while providing sufficient connectivity for content propagation. This topology better resembles Facebook-like social networks with relatively uniform degree distributions, acknowledging that true Twitter-scale dynamics require populations of tens of thousands to millions of agents where power-law distributions become statistically significant.

The network design deliberately excludes hierarchical commenting structures found in platforms like Reddit, instead implementing a flat feed structure similar to Twitter's timeline. This decision reduces computational overhead by eliminating tree-traversal operations while focusing experimental attention on content ranking mechanisms rather than conversational threading dynamics. Social connections are bidirectional, representing mutual following relationships that determine content visibility between agents.

3.2 Agent cycle

Each agent follows a structured interaction cycle consisting of content consumption, evaluation, and generation phases. During each simulation step, agents receive a personalized feed of posts from their network neighbors, ranked according to the active ranking algorithm. The agent processes each post through an LLM-powered evaluation system, making binary decisions to read-only or read-and-like based on content relevance and personal preferences encoded in their persona.

Agent memory management employs a sliding-window approach with configurable length. This bounded-memory system prevents exponential growth in computational requirements while maintaining sufficient context for coherent behavior. After consuming their ranked feed, each agent generates new content using their persona-specific instructions, contributing to the global discourse pool

for subsequent simulation steps.

The agent architecture integrates LLM APIs for both content evaluation and generation, with robust error handling and retry mechanisms to manage API limitations. Memory persistence across simulation steps enables agents to develop consistent posting patterns and social preferences, while the bounded window prevents context overflow that could degrade LLM performance.

3.3 Ranking

Our ranking system implements a modular architecture that supports multiple algorithms to study their impact on discourse dynamics. The base ranking interface combines network-level signals (global post-popularity) with individual-level preferences (personalized relevance) through weighted combination:

$$\text{score}_{u,p} = (w_{\text{network}} \cdot S_{\text{network}}(p) + w_{\text{individual}} \cdot S_{\text{individual}}(u,p))\,(1+\epsilon)$$

where ϵ represents a small multiplicative noise to break down ties, and add randomicity to the system.

We implement five distinct ranking strategies: `RandomRanker` provides baseline comparison through uniform random scoring; `LikeRanker` prioritizes posts that accumulated more likes; `UserLikeRanker` emphasizes content from historically popular users; `PersonalizedUserLikeRanker` weights posts based on individual interaction history; and `SemanticSimilarityRanker` employs cosine similarity between post embeddings and user posting history.

This ranking diversity enables systematic study of how algorithmic choices influence information propagation, echo chamber formation, and overall discourse quality in simulated social networks. The modular design facilitates easy addition of new ranking strategies for future experimental needs.

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, et al. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*, 2025. 32

Ahmed Agiza, Mohamed Mostagir, and Sherief Reda. Politune: Analyzing the impact of data selection and fine-tuning on economic and political biases in large language models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 2–12, 2024. 14, 16

Simone Alghisi, Massimo Rizzoli, Gabriel Roccabruna, Seyed Mahed Mousavi, and Giuseppe Riccardi. Should we fine-tune or rag? evaluating different techniques to adapt llms for dialogue. *arXiv preprint arXiv:2406.06399*, 2024. 13

Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. Measuring political bias in large language models: What is said and how it is said. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11142–11159, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.600. URL `https://aclanthology.org/2024.acl-long.600/`. 13

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*, 2020. 18, 32

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Simulating opinion dynamics with networks of LLM-based agents. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.211. URL `https://aclanthology.org/2024.findings-naacl.211/`. 13, 17

Lee J Cronbach. Coefficient alpha and the internal structure of tests. *psychometrika*, 16(3):297–334, 1951. 18, 31

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 8-bit optimizers via block-wise quantization. *CoRR*, abs/2110.02861, 2021. URL `https://arxiv.org/abs/2110.02861`. 33

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, Saba Sturua, Saiteja Utpala, Mathieu Ciancone, Marion Schaeffer, Gabriel Sequeira, Diganta Misra, Shreeya Dhakal, Jonathan Rystrøm, Roman Solomatin, Ömer Çağatan, Akash Kundu, Martin Bernstorff, Shitao Xiao, Akshita Sukhlecha, Bhavish Pahwa, Rafał Poświata, Kranthi Kiran GV, Shawon Ashraf, Daniel Auras, Björn Plüster, Jan Philipp Harries, Loïc Magne, Isabelle Mohr, Mariya Hendriksen, Dawei Zhu, Hippolyte Gisserot-Boukhlef, Tom Aarsen, Jan Kostkan, Konrad Wojtasik, Taemin Lee, Marek Šuppa, Crystina Zhang, Roberta Rocca, Mohammed Hamdy, Andrianos Michail, John Yang, Manuel Faysse, Aleksei Vatolin, Nandan Thakur, Manan Dey, Dipam Vasani, Pranjal Chitale, Simone Tedeschi, Nguyen Tai, Artem Snegirev, Michael Günther, Mengzhou Xia, Weijia Shi, Xing Han Lù, Jordan Clive, Gayatri Krishnakumar, Anna Maksimova, Silvan Wehrli, Maria Tikhonova, Henil Panchal, Aleksandr Abramov, Malte Ostendorff, Zheng Liu, Simon Clematide, Lester James Miranda, Alena Fenogenova, Guangyu Song, Ruqiya Bin Safi, Wen-Ding Li, Alessia Borghini, Federico Cassano, Hongjin Su, Jimmy Lin, Howard Yen, Lasse Hansen, Sara Hooker, Chenghao Xiao, Vaibhav Adlakha, Orion Weller, Siva Reddy, and Niklas Muennighoff. Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*, 2025. doi: 10.48550/arXiv.2502.13595. URL `https://arxiv.org/abs/2502.13595`. 33

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024. 33

Kevin A Fischer. Reflective linguistic programming (rlp): A stepping stone in socially-aware agi (socialagi). *arXiv preprint arXiv:2305.12647*, 2023. 13, 17

Clémentine Fourrier, Nathan Habib, Alina Lozovskaya, Konrad Szafer, and Thomas Wolf. Open llm leaderboard v2. `https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard`, 2024. 32

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024. 12, 14

Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. S3: Social-network simulation system with large language model-empowered agents. *arXiv preprint arXiv:2307.14984*, 2023. 12

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 18, 31

George Gui and Olivier Toubia. The challenge of using llms to simulate human behavior: A causal inference perspective. *arXiv preprint arXiv:2312.15524*, 2023. 12

Zhang Guozhen, Yu Zihan, Li Nian, Yu Fudan, Long Qingyue, Jin Depeng, and Li Yong. Human behavior simulation: Objectives, methodologies, and open problems. *arXiv preprint arXiv:2412.07788*, 2024. 15

Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. Binary classifier optimization for large language model alignment. *arXiv preprint arXiv:2404.04656*, 2024. 33

Maik Larooij and Petter Törnberg. Do large language models solve the problems of agent-based modeling? a critical review of generative social simulations. *arXiv preprint arXiv:2504.03274*, 2025. 12, 13, 16, 17

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in llms: Reward calibration in rlhf. *arXiv preprint arXiv:2410.09724*, 2024. 14

Junyi Li, Charith Peris, Ninareh Mehrabi, Palash Goyal, Kai-Wei Chang, Aram Galstyan, Richard Zemel, and Rahul Gupta. The steerability of large language models toward data-driven personas. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7290–7305, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.405. URL `https://aclanthology.org/2024.naacl-long.405/`. 12

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*, 2023. 17, 30, 32

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL `https://aclanthology.org/W04-1013/`. 30

Yuhan Liu, Xiuying Chen, Xiaoqing Zhang, Xing Gao, Ji Zhang, and Rui Yan. From skepticism to acceptance: simulating the attitude dynamics toward fake news. In *Proceedings of the Thirty-Third In-*

ternational Joint Conference on Artificial Intelligence, IJCAI '24, 2024. ISBN 978-1-956792-04-1. doi: 10.24963/ijcai.2024/873. URL https://doi.org/10.24963/ijcai.2024/873. 12

Yuxuan Lu, Jing Huang, Yan Han, Bennet Bei, Yaochen Xie, Dakuo Wang, Jessie Wang, and Qi He. Beyond believability: Accurate human behavior simulation with fine-tuned llms. *arXiv preprint arXiv:2503.20749*, 2025. 13, 16

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*, 2022. doi: 10.48550/ARXIV.2210.07316. URL https://arxiv.org/abs/2210.07316. 33

Simon Münker and Achim Rettinger. twony: A micro-simulation of the impact of osn mechanics on the emotionality of online discourse. 2025. 10

Simon Münker, Nils Schwager, and Achim Rettinger. Don't trust generative agents to mimic communication on social networks unless you benchmarked their empirical realism. *arXiv preprint arXiv:2506.21974*, 2025. 11

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265, 2024. 13, 17

Simon Münker. Cultural bias in large language models: Evaluating ai agents through moral questionnaires. *Proceedings of 0th Symposium on Moral and Legal AI Alignment of the IACAP/AISB Conference*, page 61–76, 2025a. 13

Simon Münker. Political bias in llms: Unaligned moral values in agent-centric simulations. *Journal for Language Technology and Computational Linguistics*, 38(2):125–138, Jul. 2025b. doi: 10.21248/jlcl.38.2025.289. URL https://jlcl.org/article/view/289. 13

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 14

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. Generative agents: Interactive simulacra of human behavior. In *In the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*, UIST '23, New York, NY, USA, 2023. Association for Computing Machinery. 13

Karl Pearson and Francis Galton. Vii. note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58(347-352):240–242, 1895. doi: 10.1098/rspl.1895.0041. URL https://royalsocietypublishing.org/doi/abs/10.1098/rspl.1895.0041. 18, 32

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023. 14, 16, 33

Luca Rettenberger, Markus Reischl, and Mark Schutera. Assessing political bias in large language models. *Journal of Computational Social Science*, 8(2):1–17, 2025. 13

Joseph Suh, Erfan Jahanparast, Suhong Moon, Minwoo Kang, and Serina Chang. Language model fine-tuning on scaled survey data for predicting distributions of public opinions. *arXiv preprint arXiv:2502.16761*, 2025. 13

V Vendetti, LD Comencini, F Deriu, V Modugno, et al. Passing the turing test in political discourse: Fine-tuning llms to mimic polarized social media comments. *arXiv preprint arXiv:2506.14645*, 2025. 13

Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, et al. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024. 13

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. In *ICLR*, 2025. 13, 17

Mozhi Zhang, Mianqiu Huang, Rundong Shi, Linsen Guo, Chong Peng, Peng Yan, Yaqian Zhou, and Xipeng Qiu. Calibrating the confidence of large language models by eliciting fidelity. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2959–2979, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.173. URL https://aclanthology.org/2024.emnlp-main.173/. 14

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025. 12

Appendix A Methods

Due to limited computational capacity, our experiments are necessarily limited in scope, utilizing a single foundation model, embedding model, and dataset. Applying the same rationale, we opt for a uniform set of hyperparameters across all experiments, also isolating the effects of our contributions from implementation-specific confounding.

A.1 Dataset

A.1.1 Discourse Reconstruction and Filtering

Our initial dataset consists of 7,790,741 tweets from 34,720 different users collected up to August 15, 2023. We identified 100 politically active seed users (users who recently engaged with politicians' content) and merged their followee networks to create the user set. For each user, we extracted their maximum available tweet history (up to 2,300 tweets per user).

We reconstruct all discourses in the dataset by identifying replies with no further reply (in the dataset) and then following the *Reply_To_ID* tag back to the original tweet. This ensures that there are no duplicates while preserving multi-branched discourse structures.

To ensure validity (primarily in the test set), we opt for a minimal pre-processing principle. However, we exclude conversations where the final reply contains URLs (typically image links, news references, or retweets) to prevent models from learning to generate URL patterns rather than meaningful discourse content. Additionally, we filter out conversations containing messages with both URLs and fewer than 10 words, since such cases typically depend on URL content for context that cannot be reconstructed from the text alone. For training efficiency, we removed length outliers, specifically conversations where the final reply length fell below the 10th or above the 90th percentile, as well as any conversation containing a message exceeding the 90th percentile threshold.

A.1.2 Chronological Demonstration Format

Our approach to prompt construction differs fundamentally from standard user simulation methods, where handcrafted descriptions of persona characteristics guide the generation. Instead, we group all conversations by the user who authored the final reply, creating user-specific demonstration sets. Each prompt consists of a minimal system instruction ("*You are a social media user responding to conversations. Keep your replies consistent with your previous writing style and the perspectives you have expressed earlier.*") followed by n-shot demonstrations of the target user's previous conversations in chronological order (Figure 1).

For each user with more than $n + 1$ conversations, we implement a sliding-window approach to maximize training sample efficiency. This method creates multiple training instances per user by shifting the demonstration window forward one conversation at a time while preserving temporal ordering.

For example, with a 3-shot configuration, conversations 1-3 serve as demonstrations for predicting conversation 4, then conversations 2-4 for predicting conversation 5, and so on. This approach preserves chronological integrity while increasing the available training samples.

The prompt structure follows a standardized format where each demonstration consists of a complete discourse exchange (original post, intermediate replies, and the user's response), with the final prompt presenting only the conversation context without the target user's reply. This methodology allows the model to learn from authentic user behavior patterns without explicitly encoding assumptions about which characteristics influence response generation.

To clearly signal what each user wrote - especially in multi-turn conversations - we prefix each message with a greater-than sign followed by the username and content, as shown in Prompt 2:

```
>{username1}: {Tweet}
>{username2}: {Reply1}
```

Prompt 2: Intra Prompt Conversation Format

When a user refers to the person being simulated (the LLM), we replace the actual username with *@YOU*. If the user being simulated has already replied in a previous stage of the conversation, the username is replaced by *>YOU:*. The assistant component of each prompt consists solely of the target reply.

These conversation-reply pairs are tokenized using LLM-specific chat templates. Since these models are trained on human-assistant dialogue patterns, the conversation serves as the user input while the reply represents the model's expected output. This process conditions the LLM by presenting the dialogue history as if it had already generated those responses during prior turns.

A.1.3 Cross-User Evaluation Design

The sliding window approach leads to an imbalanced distribution in which a small number of "power users" contribute disproportionately to the training samples. To prevent the model from primarily learning patterns from these frequent users, we first remove all users above the 90th percentile in conversation frequency, as these extremely active users would otherwise dominate the training distribution and potentially skew model behavior toward their specific discourse patterns.

We then bin the remaining users according to their conversation frequency using density-based

clustering to create strata of similarly active users. From each stratum, we randomly assign 15% of the users to the test set, ensuring that the test distribution maintains the same proportion of user activity levels as the training set.

This user-based split methodology has two key advantages: it maintains ecological validity by preserving realistic activity distributions in both training and test sets, and it directly tests the model's ability to extract generalizable discourse patterns rather than memorizing user-specific idiosyncrasies.

For our experimental framework, we further divide the training data into two equal parts. The first part serves for initial SFT adaptation, while the second part is used exclusively for preference learning, making the approach less prone to overfitting and encouraging robust generalization.

A.1.4 Dataset Size and Construction

The dataset construction process generated varying amounts of examples in different n-shot configurations:

- 3-shot configuration: 13,852 training examples, 1,000 evaluation examples
- 5-shot configuration: 9,067 training examples, 1,000 evaluation examples
- 7-shot configuration: 5,984 training examples, 690 evaluation examples

A.2 Metrics

A.2.1 Embedding Distance

We measure the semantic similarity between the responses generated by the model and the actual human responses through the cosine distance between their embedding representations, using *gte-Qwen2-1.5B-instruct* with discourse-specific prompting (Prompt 1) (Li et al., 2023). This metric captures alignment in both communicative intent and stylistic characteristics, with lower distances (ranging from 0 to 2) indicating closer approximation of the human user's response pattern.

A.2.2 ROUGE-1

We employ ROUGE-1 (unigram overlap) to quantify the lexical similarity between the generated and actual human responses (Lin, 2004). This surface-level metric captures the model's adoption of user-specific vocabulary, hashtag usage, and other explicit linguistic features common in social media discourse. We implement the metric using the rouge-score library and report F1 scores.

A.2.3 Perplexity

We calculate token-level perplexity as an information-theoretic measure of how well each model predicts the actual human responses given the conversation history. Perplexity directly measures the model's probability assignment to ground-truth responses, providing insight into its internal alignment with human discourse patterns. Lower perplexity indicates that the model assigns a higher probability to the tokens actually produced by users. Our implementation creates a custom attention mask that isolates the calculation to only the target response, preventing context tokens from influencing the measurement.

A.2.4 Political Stance Consistency

To assess the political stance modeling capabilities we sample 30 random histories in the 5-shot format. We then append a made-up conversation asking for the "take" on one of the following politically charged topics: abortion, healthcare, LGBTQ rights, gun rights, immigration, and climate change, as shown in Prompt 3:

```
>NilsSchwager: What's your take on {topic}?
```

Prompt 3: Probe for Political Stance on Controversial Topics

This pushes the model to generate a reply that reveals a political leaning. To assess the political leaning, we insert the reply and the tweet (giving the LLM-as-a-Judge additional context) in the prompt, as shown in Prompt 4. The LLM-as-a-Judge is tasked to determine the political leaning on a scale from (1) strongly conservative to (5) strongly progressive/liberal. By enforcing structured output we guarantee the output is automatically parsable.

We opt for *Llama 3.1 70B* (Grattafiori et al., 2024) as LLM-as-a-Judge model. We prompt the model to compare the responses to typical conservative/liberal positions for its assessment, relying totally on the model's calibration. While there are more nuanced approaches, we would argue that in our case this is the best approach, because we focus solely on the consistency across topics. Given the construct *political leaning* is stable across the diverse topics within the judge we deem our approach sufficient.

For each tested model we generate 10 replies and average them for each topic-history combination. This results in 30 profiles (histories) with 6 scores (topics). We would expect that a well calibrated model detects a political leaning in the history and is able to reproduce it across the other topics, resulting in constant scores and corresponding high value for Cronbach's Alpha (Cronbach, 1951).

```
Analyze this social media reply in the context of the conversation:
Original conversation: {last_conversation_content}

Reply: {synthetic_reply}

Evaluate the ideological leaning of this reply regarding
{target_topic} on a scale from 1-5:

1 = Strongly conservative position (supporting conservative
    policies or opposing progressive policies)
2 = Moderately conservative position
3 = Neutral, moderate, or mixed ideological stance
4 = Moderately progressive/liberal position
5 = Strongly progressive/liberal position (supporting
    progressive policies or opposing conservative policies)

Focus specifically on the underlying political ideology, not
writing style or tone. Consider how the views expressed align
with typical conservative versus progressive positions on
{target_topic}.

Please provide your evaluation in the following format (JSON-like):
"rating": "1|2|3|4|5"
```

Prompt 4: LLM-as-a-Judge Prompt for Political Stance Evaluation

A.2.5 Offensiveness Correlation

Using the TweetEval offensive content classifier, we analyze the correlation between offensive content in conversation histories and corresponding model-generated responses (Barbieri et al., 2020). For each history we generate the reply ten times and average the ten corresponding offensiveness scores. By averaging across ten runs we get a more robust score for each setting. We also calculate the offensiveness score from the replies in the history (the style the model is tasked to replicate) and then calculate Pearson correlation coefficients (Pearson and Galton, 1895). This quantifies each model's ability to reproduce varying levels of sensitive content.

A.3 Model Selection

We employ Microsoft's *Phi-4-mini-instruct* (3.8B parameters) as our foundation model for all experiments (Abouelenin et al., 2025). This model achieves superior performance on the Open LLM Leaderboard[1] among all models under 7B parameters (Fourrier et al., 2024). The parameter count constraint was necessary to enable computationally efficient training and evaluation within our resource limitations. Applying the same rationale, we opt for *gte-Qwen2-1.5B-instruct* (Li et al., 2023) as our embedding

[1] as of 29.04.2025

model, which ranks second in the Massive Text Embedding Benchmark (Enevoldsen et al., 2025; Muennighoff et al., 2022) among all models under 7B parameters that offer instruction retrieval [2].

A.4 Hyperparameters

We employ full-parameter training with *bfloat16* precision across all experiments, foregoing parameter-efficient fine-tuning methods like LoRA or QLoRA to eliminate additional hyperparameter complexity. All fine-tuning methods utilize the memory-efficient *paged_adamw_8bit* optimizer (Dettmers et al., 2021) with consistent configurations. The maximum sequence lengths are adjusted proportionally to the demonstration quantity: 512 tokens for 3-shot, 768 for 5-shot, and 1024 for 7-shot datasets.

For SFT, we used a learning rate of 2e-5 with a batch size of 8 and linear warm-up over 10% of the training steps. All preference-based approaches (DPO, CPO, BCO, KTO) use a reduced learning rate of 1e-6, maintaining a standardized effective batch size of 16 across methods through appropriate gradient accumulation. Each preference learning approach trains for 3 epochs. These choices result from our hyperparameter tuning comparing epochs and commonalities in the literature (Ethayarajh et al., 2024; Rafailov et al., 2023; Jung et al., 2024).

[2] as of 29.04.2025 - leaving out multilingual-e5-large-instruct

Appendix B Results

Approach	3-Shot	5-Shot	7-Shot
Base Model	0.5235 (± 0.0042)	0.4767 (± 0.003)	0.45 (± 0.0038)
Initial Fine-Tuning on Authentic Completions			
SFT Completion Only	0.3935 (± 0.0031)	0.3895 (± 0.004)	0.3919 (± 0.0026)
SFT Full-Context	0.3906 (± 0.003)	0.3808 (± 0.0027)	0.3777 (± 0.0028)
Fine-Tuning on Synthetic Completions			
SFT+BCO	0.3796 (± 0.0029)	0.377 (± 0.0031)	0.3776 (± 0.0048)
SFT+CPO	0.3868 (± 0.0026)	0.3782 (± 0.0038)	0.3847 (± 0.0047)
SFT+DPO	**0.3771** (± 0.0023)	**0.3731** (± 0.0031)	0.3765 (± 0.0053)
SFT+KTO	0.3806 (± 0.002)	0.3763 (± 0.0039)	**0.3749** (± 0.0027)
SFT+SFT	0.3791 (± 0.0041)	0.3756 (± 0.0038)	0.3841 (± 0.0033)
Fine-Tuning on Authentic Completions (control)			
SFT+BCO	0.3919 (± 0.0028)	0.3859 (± 0.0026)	0.3908 (± 0.0047)
SFT+CPO	0.3964 (± 0.0038)	0.3881 (± 0.0027)	0.3913 (± 0.0047)
SFT+DPO	0.3925 (± 0.0041)	0.3861 (± 0.0052)	0.3898 (± 0.0043)
SFT+KTO	0.3925 (± 0.0026)	0.3863 (± 0.0029)	0.3872 (± 0.004)
SFT+SFT	0.389 (± 0.0024)	0.3795 (± 0.0034)	0.3784 (± 0.0049)

Cosine Distance Comparison across Fine-Tuning Methods (Lower Values Indicate Better Performance). Standard deviations shown in parentheses. Results averaged across 10 independent runs: 3-shot and 5-shot configurations based on 1000 test samples; 7-shot based on 690 samples. Best performance in each column is highlighted in **bold.**

Approach	3-Shot	5-Shot	7-Shot
Base Model	0.1287 (± 0.001)	0.1373 (± 0.0011)	0.1438 (± 0.0021)
Initial Fine-Tuning on Authentic Completions			
SFT Completion Only	0.1987 (± 0.0015)	0.1996 (± 0.0028)	0.1943 (± 0.0022)
SFT Full-Context	0.1985 (± 0.0022)	0.2009 (± 0.0016)	**0.1998** (± 0.0025)
Fine-Tuning on Synthetic Completions			
SFT+BCO	0.2019 (± 0.0027)	0.2005 (± 0.0023)	0.1961 (± 0.0022)
SFT+CPO	0.1955 (± 0.0026)	0.197 (± 0.0027)	0.1939 (± 0.0028)
SFT+DPO	0.2038 (± 0.0019)	0.1956 (± 0.0021)	0.197 (± 0.0034)
SFT+KTO	0.2044 (± 0.0025)	0.2006 (± 0.0025)	0.1978 (± 0.0033)
SFT+SFT	**0.2075** (± 0.0023)	**0.2032** (± 0.0029)	0.1959 (± 0.0018)
Fine-Tuning on Authentic Completions (control)			
SFT+BCO	0.1956 (± 0.0018)	0.1973 (± 0.0021)	0.1928 (± 0.0027)
SFT+CPO	0.1915 (± 0.0019)	0.1936 (± 0.003)	0.1926 (± 0.0029)
SFT+DPO	0.1952 (± 0.0043)	0.1976 (± 0.002)	0.1951 (± 0.0028)
SFT+KTO	0.1944 (± 0.0037)	0.1986 (± 0.0027)	0.1939 (± 0.0029)
SFT+SFT	0.2001 (± 0.0031)	0.1977 (± 0.0014)	0.1984 (± 0.0015)

ROUGE-1 Comparison across Fine-Tuning Methods (Higher Values Indicate Better Performance). Standard deviations shown in parentheses. Results averaged across 10 independent runs: 3-shot and 5-shot configurations based on 1000 test samples; 7-shot based on 690 samples. Best performance in each column is highlighted in **bold.**

Approach	3-Shot	5-Shot	7-Shot
Base Model	5e15	7.9e14	4.04e12
Initial Fine-Tuning on Authentic Completions			
SFT Completion Only	14.3111	14.3864	13.8682
SFT Full-Context	50.9115	51.9867	56.8716
Fine-Tuning on Synthetic Completions			
SFT+BCO	15.2194	14.3925	13.9061
SFT+CPO	14.4378	14.4689	13.9335
SFT+DPO	14.4395	14.4338	13.8997
SFT+KTO	18.2191	18.5141	13.9058
SFT+SFT	17.072	16.0201	15.2657
Fine-Tuning on Authentic Completions (control)			
SFT+BCO	14.2681	14.3303	13.815
SFT+CPO	14.3156	14.3938	13.8687
SFT+DPO	14.2573	14.335	13.8311
SFT+KTO	14.2519	14.3422	13.825
SFT+SFT	**13.9889**	**13.9663**	**13.5245**

Perplexity Results across Fine-Tuning Methods (Lower Values Indicate Better Performance). 3-shot and 5-shot configurations based on 1000 test samples; 7-shot based on 690 samples. Best performance in each column is highlighted in **bold.**

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