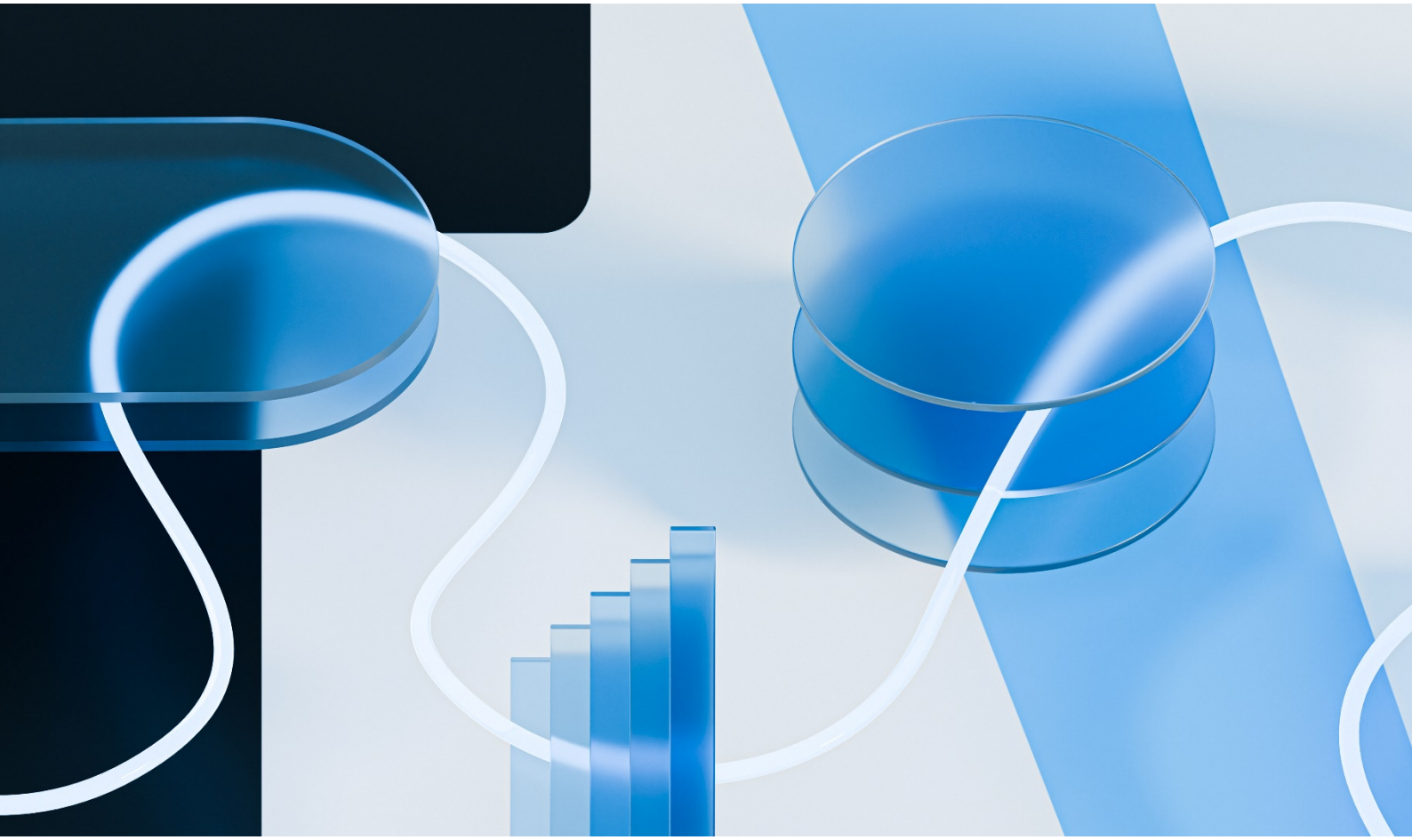
twon
twin of online social networks



TWin of Online Social Networks

Deliverable D3.1

Deep Learning Report

Main Authors: Simon Münker & Achim Rettinger

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	D3.1
Deliverable Name	Deep Learning Report
Due Date	31.03.2025
Submission Date	XX.03.2025
Type	R — Document, report
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

Executive Summary

The TWON (Twin of Online Social Networks) project aims to develop a comprehensive understanding of online social networks (OSNs) through the creation of realistic simulations. This report, Deliverable D3.1, focuses on the application of large language models (LLMs) to simulate user behavior within OSNs, with a particular emphasis on text-centered platforms. The research is structured around Work Package 3 (WP3), which breaks down the complex architecture of social networks into manageable subcomponents, focusing on the development of generative agents that can effectively mimic human interactions.

Key Contributions and Findings

Initial LLM Assessment The report begins with an evaluation of various LLMs to determine their suitability for simulating social media interactions. ChatGPT emerged as the most capable model for generating authentic social media content, while open-source models like Llama-2-70B and Falcon-180B showed limitations. This assessment led to the development of the Agent Stylized Social Actions (A2SA) methodology, which provides a flexible prompt-based pipeline for modeling user interactions.

Multi-Language Persona Experiment A preliminary experiment evaluated LLM performance across multiple languages (English, German, Dutch) and political orientations (liberal, conservative, alt-right). The results revealed significant variations in authenticity scores across languages, with English content achieving the highest quality metrics. This highlighted the challenges of simulating authentic discourse in non-English contexts, particularly for low-resource languages like Dutch.

Moral Foundations Assessment The section investigates the capacity of LLMs to represent specific political ideologies through moral foundation questionnaires. The findings indicate that the models show a response inconsistency. Also, most models exhibited a measurable left-leaning bias. This raises important questions about the reliability of using LLMs as proxies for human belief systems in social science research.

Data-Driven Alignment Framework The most substantial methodological contribution is the development of a formal framework for building TWONs through data-driven agent behavior modeling. This framework separates content creation (posting) from reaction (replying, liking) and implements specialized machine learning methods for each task type. Fine-tuning approaches, particularly for English content, showed significant improvements in empirical realism scores.

twon
twin of online social networks

Contents

List of Tables

List of Figures

List of Listings

List of Abbreviations

BERT	Bidirectional Encoder Representations from Transformers
LLM	Large Language Model
MFQ	Moral Foundation Questionnary
OSN	Online Social Network
TWON	Twin of Online Social Networks

Deep Learning Report

Simon Münker & Achim Rettinger *

March 21, 2025

1 Introduction

The rapid evolution of online social networks (OSNs) has fundamentally transformed the way individuals interact, share information, and form opinions. As these platforms continue to grow in complexity and influence, understanding the underlying mechanisms that drive user behavior and content dissemination has become a critical area of research. The ability to model and simulate these interactions not only provides insights into the dynamics of social networks but also offers a powerful tool for predicting and mitigating potential risks, such as the spread of misinformation, polarization, and harmful content. This report presents a comprehensive exploration of the use of large language models (LLMs) in simulating and analyzing user behavior within OSNs, with a particular focus on text-centered platforms like Twitter (now X). The work builds on the foundational progress initiated during the TWON Hackathon in Karlsruhe, where the technical implementation of machine-learned aspects of social networks was first conceptualized. The project, structured around Work Package 3 (WP3), aims to break down the complex architecture of social networks into manageable subcomponents, focusing on the development of generative agents that can effectively mimic human interactions. Through a series of experiments and methodological innovations, this report addresses critical questions about the feasibility and limitations of using LLMs to simulate realistic social media behavior. We evaluate various language model architectures, explore their ability to represent diverse viewpoints, develop a formal framework for quantifying empirical realism, and assess the unique challenges that arise when applying these technologies to multilingual European contexts. The report concludes with a discussion of both the technical and ethical considerations that should guide future research in this rapidly evolving field.

*This report draws strongly on collaborative research projects. We would like to acknowledge the co-authors of these papers: Nils Schwager (UT), Kai Kugler (UT), Michael Heseltine (UvA), and Sjoerd Stolwijk (UvA). We would like to thank Damian Trilling (UvA) for reviewing the internal draft of the paper. The report has benefited considerably from the comments.

1.1 Integration with the TWON Project

1.1.1 Contribution to the Overall Project

Our workpackage serves as a critical bridge between the theoretical foundations developed in WP2 (Computational Modelling of Complex Social Networks) and the practical implementation in WP4 (TWON Implementation and Computation). By providing data-driven estimation techniques for TWONs, we ensure that the abstract models can be accurately calibrated using real-world observations, thus significantly enhancing the empirical validity of the entire TWON approach.

The machine learning models we develop directly contribute to Obj1 (Build TWONs of OSNs) by enabling parameter estimation for the complex network models, and indirectly support Obj3 (Simulate Democratic Debates) by providing realistic parameterizations that allow for meaningful counterfactual simulations. This positions our workpackage as an essential component in the methodology outlined in Sec. 1.2.1.3, where the co-evolution of theory-driven and data-driven approaches is central to the project's success.

1.1.2 Dependencies on Other Workpackages

Our workpackage relies on several inputs from other components of the TWON project:

- **From WP2:** We require the formal specifications of model parameters and differentiable functions that capture the essential dynamics of OSNs. The success of our machine learning approach is contingent on how well these models can represent real-world phenomena while remaining amenable to optimization techniques.
- **From WP4:** We depend on the data acquisition and processing infrastructure to obtain the four levels of data (DS1-DS4) described in Sec. 1.2.1.4. Particularly, the alignment between these different data sources is crucial for the success of our parameter estimation task.
- **From WP5:** The case studies and experimental designs provide the contextual framework and ground truth data against which our estimations can be validated, particularly relevant for the iterative refinement of our models during the two project cycles.

1.1.3 Outputs Used by Other Workpackages

The outputs from our workpackage serve as critical inputs for multiple other components:

- **For WP2:** Our parameter estimations provide empirical validation for theoretical models, allowing researchers to refine their assumptions and focus on the most relevant mechanisms in subsequent iterations.
- **For WP4:** The machine-learned model parameterizations are directly incorporated into the TWON implementation, enabling realistic simulations of OSN dynamics that closely match observed behavior.
- **For WP5:** Our calibrated models enhance the ecological validity of the field studies by ensuring that the experimental TWON platforms used by participants accurately reflect the dynamics of real-world OSNs.
- **For WP6:** The insights generated from our parameter estimations contribute to evidence-based recommendations, particularly regarding how specific platform mechanics influence democratic debates.

Through these bidirectional relationships with other workpackages, our work constitutes a central node in the TWON methodology, translating between abstract theory and concrete implementation while ensuring that the entire project maintains strong empirical grounding throughout the research process.

1.2 Research Components and Contributions

1.2.1 Initial LLM Assessment

Our preliminary investigation evaluated various LLMs for their capability to generate authentic social media content. We systematically compared publicly available models (ChatGPT, Llama-2-70B, Falcon-180B, Luminous-Supreme) alongside locally deployed smaller models within our hardware constraints. The results demonstrated that ChatGPT consistently produced the most authentic content, exhibiting appropriate stylistic features for Twitter discourse, including hashtags and emojis. Even the largest open-source models showed limitations in generating context-appropriate replies, while local models with our available 80GB VRAM performed inadequately for the required tasks. This assessment led to the development of our Agent Stylized Social Actions (A2SA) approach, which provides a versatile prompt-based pipeline adaptable to both high-performance third-party models and fine-tuned smaller models. The A2SA methodology facilitates two primary interaction types: choice-based actions (reading, liking, sharing) and generative tasks (replying), both essential for modeling realistic social network behavior.

1.2.2 Multi-Language Persona Experiment

The Ettmaal Conference experiment advanced our understanding of model performance across multiple dimensions: languages (English, German, Dutch), political orientations (liberal, conservative, alt-right), and discourse metrics (topic alignment, persona authenticity). This experiment confirmed that models like GPT-3.5 and Mistral-7B can generate coherent, topic-aligned content but revealed significant variations in authenticity scores across languages. English content consistently achieved the highest quality metrics (authenticity: 4.089 for GPT-3.5), followed by German (3.900), with Dutch exhibiting substantially lower perceived authenticity (2.820) due to incorrect phrasing and American-centric perspectives. These findings highlighted important considerations for the project's European focus, particularly regarding the challenge of simulating authentic discourse in non-English contexts. The observed language disparities closely align with recent research on linguistic limitations in LLMs (Ma et al., 2024), which has remained a consistent challenge throughout the project timespan despite ongoing improvements in model capabilities.

1.2.3 Moral Foundations Assessment

Our investigation into LLMs' capacity to represent specific political ideologies through moral foundation questionnaires (Graham et al., 2009) provided critical insights into the limitations of using these models as proxies for human belief systems. By systematically prompting models with varying political personas (liberal, moderate, conservative) and measuring their moral foundation scores, we identified significant inconsistencies in response patterns across different models and personas. Notably, models like Mixtral-8x7b exhibited the highest response consistency (variance: 0.030), while others like Qwen-72b showed 14 times higher variance (0.425). Most models demonstrated a measurable left-leaning bias, aligning better with liberal human participants than conservative ones—a finding consistent with other recent research on political biases in language models (Rozado, 2023; Rutinowski et al., 2024). This component critically contributes to TWON by establishing empirical boundaries for the degree to which LLMs can realistically represent diverse viewpoints within simulated networks—essential knowledge for accurately modeling political discourse and polarization dynamics.

1.2.4 Data-Driven Alignment Framework

The most substantial methodological contribution is our formal framework for building twins of online social networks through data-driven agent behavior modeling. This approach separates content creation (posting) from reaction (replying, liking) and implements specific machine learning methods

optimized for each task type. Our experiments with fine-tuning Llama-3.2-3B produced significant improvements over basic prompting approaches, particularly for English content, with fine-tuned reply generation achieving remarkably high empirical realism scores (BLEU: 0.734, unigram precision: 0.782). This framework directly addresses the core technical requirements for TWON by providing quantifiable metrics for empirical realism and establishing benchmarks for future simulation development. The successful demonstration of fine-tuning approaches also coincides with broader advancements in LLMs during the project period, particularly the emergence of more efficient adaptation techniques like LoRA (Hu et al., 2021) that enable effective model specialization without excessive computational requirements.

1.3 Limitations and Deviations

Despite the progress achieved, several challenges emerged during implementation that necessitated adjustments to our original research plan:

- **Data Quality and Availability:** The proposal assumed access to high-quality, aligned datasets for model training. However, data provided by UvA contained corrupted identifiers, preventing full utilization for our proposed data alignment approach. According to the proposal, JSI was responsible for providing additional data (WP4, T4.1), but the data acquisition process has proven more challenging than anticipated, limiting our ability to create comprehensive simulation models with the desired granularity of user types.
- **Computational Infrastructure Constraints:** The lack of sufficient GPU resources at the University of Trier (UT) restricted our ability to utilize larger models or conduct extensive fine-tuning experiments. While the LLM field has seen remarkable progress in model efficiency during the project timespan, the computational requirements for state-of-the-art performance still exceed our available infrastructure, necessitating adaptations to work within these constraints.
- **Cross-Lingual Performance Challenges:** Our initial plan assumed comparable performance across languages, but experiments revealed significant challenges with non-English content generation. The proposed simulation in Serbian (part of the initial research proposal) appears particularly challenging given the observed performance degradation even for medium-resource languages like German. This limitation reflects broader issues in the multilingual capabilities of current LLMs, which continue to show English-centric performance patterns despite ongoing research efforts.

- **Scope Limitations:** Our current implementation focuses primarily on single-turn interactions rather than extended conversation threads. While our framework theoretically supports multi-turn discourse analysis, the empirical evaluation has concentrated on one-turn communicative behavior, leaving more sophisticated discourse metrics and longer conversational simulations for future work.

These limitations highlight important considerations for the ongoing development of TWON and point to specific areas where additional resources and methodological innovations are needed. They also reflect the rapidly evolving nature of LLM research, where capabilities and limitations continue to shift as new models and techniques emerge.

2 Initial Proposal

Research Insight

Having established the overall objectives of the TWON project, our research began with a systematic evaluation of various large language models to determine their suitability for simulating social media interactions. This assessment was deemed a necessary first step to identify which models could reliably generate authentic social media content within our technical constraints, thereby establishing the technological foundation for subsequent experiments. The investigation focused on comparing both commercial and open-source models across different parameter sizes to establish a performance baseline for our agent-based simulation approach.

To effectively model online social networks (OSN), we need to generate a user base that interacts with the messages they receive. We focus our work on text-centered OSNs with a restricted scope of user actions. These are divided into two categories. The first interaction is choice-based: reading, liking, or sharing content, and the second is generating custom content: writing a reply. We suggest large language models (LLM) to model the agents as they show strong capabilities in generating text in different styles and are able to handle multi-tasking. During our initial assessment, we focus on evaluating different approaches to model generative agents with LLMs and develop, based on our findings, a suitable proposal for implementation. During the work, we utilized the expert knowledge of computational linguistics at the University of Trier (WP3) and formulate a strong baseline covering a variety of approaches during our preliminary experiment. To narrow down an appropriate solution for implementing generative agents, we discussed and tested a set of common approaches in the domain of LLMs. We examined

a classical single task-centered architecture and tested the inferencing capabilities of modern LLMs on local and third-party provider scope.

2.1 Inferencing a Manufactured Multi-Task LLM

To align our research with trends in Natural Language Processing (NLP), we conducted experiments with modern LLMs, both locally and provided by third parties. We chose instruction fined-tuned LLMs that operate through a text-to-text interface. The models have the benefit that they can generalize through a large amount of pre-training data and can handle multiple tasks through diverse fine-tuning in a diverse range of downstream applications.

ChatGPT (OpenAI): As our baseline, we chose ChatGPT. It is one of the best-performing LLMs and utilized in the paper Generative Agents: Interactive Simulacra of Human Behavior by Park et al. (2023). The preceding work utilizes the model to generate computational software agents that simulate human behavior in a restricted environment. The simulation produces believable individual and emergent social behaviors. We advise following the website that showcases the project as a webpage. Our exemplary inferencing confirms that ChatGPT produces authentic content, and our results show an appropriate language style for discourses on Twitter, including the correct use of hashtags and emojis.

Llama-2-70B (HuggingChat): Llama-2 is the current state-of-the-art LLM trained by Meta AI. It outperforms previously released open-source language models on established benchmarks. Our results show coherent replies to the inputs. From our perspective, the content is comparable to ChatGPT but slightly less varied. The model uses fewer hashtags and emojis and a more rigid sentence structure.

Falcon-180B (HuggingChat) : During the end of the research period, the Technology Innovation Institute - Abu Dhabi released the largest version of their Falcon version. The new model supposedly surpasses Llama-2. However, we found no extensive benchmarks that prove this claim. We see less varied content and fewer emojis compared with Llama-2 in our experiments.

Luminous-Supreme (AlephAlpha): To include an EU-based model, we added Luminous to our experiment. From a legal perspective, using a model compliant with EU data privacy guidelines seems appropriate. However, the model does not perform as expected with our provided prompts. We restructured the prompts to allow for a natural completion instead of instruction formulation. Even with the adaption and further hyperparameter optimization, the model does not respond with coherent replies aligned with our inputs. The quality of the results is not comparable to ChatGPT, Llama-2, and Falcon.

[Flan|Alpaca]-T5-XXL (local): All previously introduced models are hosted and queried using a third-party provider. That limits our ability to modify and analyze the results on a machine-learnable level. To determine the possibilities of the largest local models for inferencing with respect to our computation capacities, we prompted two more models. Both show incoherent and unusable responses unaligned to our prompts.

We conclude from our preliminary experiments that the generation of agent-specific replies is possible given a powerful enough LLM. However, only ChatGPT provides us with convincing results. Even two of the largest and best open-source models struggle to generate genuine replies based on a simple agent description. Further, local models – with the restrictions to 80GB VRAM – do not perform appropriately for our task. Based on the results, we envision two possible scenarios. We utilize a third-party LLM (ChatGPT) or adapt a smaller version of a mediocre-performing model (Falcon, Llama-2) to our task. In the following, we provide a method that can be applied to both cases.

2.2 Agent Stylized Social Actions (A2SA)

A social media agent can perform a variety of activities on a platform. We propose a flexible approach, utilizing a text-to-text prompt-based pipeline based on a single LLM. Thus, we can extend and modify our actions during the project and adapt them to new platforms. Our prompts include examples (few-shot) to enable in-context learning. Depending on the task, we provide three positive and negative examples for classification or six style examples for content generation. Our proposed approach aligns with state-of-the-art research in NLP and LLMs.

Read | Like | Share (Binary Classification) Using social media, a user or agent is frequently confronted with a choice: He can either read a post or ignore it. This choice extends to the action of liking or sharing. Our first prompt template handles those actions by providing positive and negative examples of the agent's past behavior and the current action and content he faces. Our generic template allows for easy adaptation to new binary action types.

Reply (Text Generation) The second kind of interaction we model is the generation of custom content in the scope of our work restricted to text. We reframe the task of producing agent-specific replies to replicate the style of example inputs. Our approach is based on research showing the strong performance of LLMs in different writing styles. However, we note that using style as a guideline may impair our ability to manipulate the contents in defined socio-demographic dimensions and, thus, reduce the sociological insights gained.

```
operator: Literal['read', 'like', 'share']

x: str = f'''
A social media user does {operator} the following content on the plattform:
    """{user_sample_postive_1}"""
    """{user_sample_postive_2}"""
    """{user_sample_postive_3}"""

In constrast, he does not {operator} the following content:
    """{user_sample_negative_1}"""
    """{user_sample_negative_2}"""
    """{user_sample_negative_3}"""

Would the descibed user {operator} the following content:
    """{user_train_input}"""
'''

y: bool = f'{user_gold_action}'
```

Listing 1: Examplary Decision Prompt

```
x: str = f'''
A social media user reacts in the following style to content he read on the platform:
    """{user_sample_reply_1}"""
    """{user_sample_reply_2}"""
    """{user_sample_reply_3}"""
    """{user_sample_reply_4}"""
    """{user_sample_reply_5}"""
    """{user_sample_reply_6}"""

Reply in the style of the described user to the following input:
    """{user_train_input}"""
'''

y: str = f'{user_gold_reply}'
```

Listing 2: Generation prompt example

Using the proposed prompts and the appropriate raw data, we can create an aligned and enriched dataset representing the D3.2 of our work plan. The dataset can be used in two ways:

Pure Inference As our preliminary results show, ChatGPT is capable of producing authentic content without examples (zero-shot). We expect, backed up by current research, that with few shot-scenario shows further improvements (see Sec. 5). The paper Generative Agents: Interactive Simulacra of Human Behavior by Park et al. (2023) demonstrates the viability of this approach in combination with ChatGPT. Therefore we estimate a high chance of success for this approach. However, relying on an external provider may not align with the goals of the TWON research project and result in additional costs. We suggest discussing this during the consortium meeting to steer the subsequent research in the desired direction.

Fine-Tuning If we can create a sufficiently large dataset, we can fine-tune a pre-trained LLM to follow our specific instructions. With this approach, we have control over our model and can specifically modify it to our use case. In contrast to a pure inference approach, our fine-tuned LLM aim to fulfill the nees of D3.3 (Prototype of calibrated TWON with machine learned parametrization). However, we classify the feasibility lower in contrast to the first approach, as it heavily relies on the quantity/quality of the data and the fine-tuning process. Both are potential challenges during implementation.

Parameter Efficient Fine Tuning: To utilize the largest possible LLM without sacrificing training efficiency, state-of-the-art approaches only adapt a selection of the model parameters or add trainable layers between architectural components.

2.3 Challenges during Implementation

Data Quality/Quantity: UvA provided the first batch of data. Due to a corruption of the identifiers, we cannot fully utilize the data for our proposed data alignment. Thus, we need more data in an appropriate format. According to the proposal, JSI is responsible for providing the data (WP4, T4.1). The data question is an imminent issue that we need to prioritize in the upcoming project months, regardless of the decision concerning the neural agent modeling.

Research Insight

Following these experiments, it became clear that while high-performance third-party models like ChatGPT offered superior performance, their integration into our research pipeline presented both practical and ethical challenges. These findings directly informed our subsequent development of the A2SA methodology and highlighted the need for fine-tuning smaller models as an alternative approach. This initial assessment thus provided critical direction for the project's technical implementation strategy and established key performance benchmarks against which future improvements could be measured.

3 Preliminary Experiment

Research Insight

Building upon our initial model assessments, the Ettmaal Conference experiment was designed to evaluate LLM performance across the multiple dimensions required for realistic simulation of European social media discourse. This experiment represented a critical step in determining whether current models could adequately capture the linguistic and cultural nuances necessary for the TWON project's cross-national scope. By systematically varying languages and political orientations, we sought to identify potential limitations in model capabilities that might affect the ecological validity of our simulations.

3.1 Methods

Our experiments depend on the selection of language models and the ensemble of our textual input as a base for the generated content. In the following, we clarify the relevant aspects and connect them to your dimensions and analysis interests.

Language Models In our current experiment iteration, we compare two language models differing in the number of parameters. As an established gold standard for text generation (Wu et al., 2023), we use GPT3.5 by Open AI with 175B parameters. The model is the foundation for ChatGPT and extended the reputation of language models beyond the domain of computational linguistics. Comparable studies (Törnberg et al., 2023) verify that the model can generate authentic social media content. As a second model, we choose Mistral-Instruct by Mistral AI, a European company providing open-access solutions for language models. In contrast to GPT3.5, the model has 7B parameters, 25 times smaller. Mistral-

Instruct, however, shows remarkable performance, comparable to larger models, in established text generation benchmarks (Jiang et al., 2023).

Recent studies (Li et al., 2023) suggest that smaller language models can generally perform equivalently or even outperform larger ones. These promising results make Mistral-Instruct worth investigating for social media content generation. Focusing on small models is practically motivated. The improved efficiency, lower cost, and easy deployment are beneficial beyond our studies for future work on language model usage for studying social networks.

Prompting The chosen models stand on the text-2-text paradigm (Raffel et al., 2020). Thus, they rely on text as input and return text as output. It allows solving tasks – like classification and generation – while providing only a textual description, a prompt. Thus, optimization of the model turns from adapting parameters to tuning words used as instructions. While this paradigm shift heavily impacts the workflow of machine learning applications and shifts the research focus of natural language processing with language models, our experiments focus on the generated content and not the optimization of the prompts.

We utilize a zero-shot prompting approach (Kojima et al., 2022), meaning an instruction without examples. We explicitly state that the model has no preceding interaction with the platform. The prompt asks the model to generate a social media post based on the following interchangeable variables. They reflect your analysis dimensions for the perceived quality of generated content.

Language We generate content for three languages: English, German, and Dutch. Both models are primarily trained on English corpora, with less training data seen in German and even fewer in Dutch. On the technical side, it allows us to analyze the capability of the models to generalize across languages and their multilingual performance. Content-wise, it enables us to determine to what degree the models adapt typical views of the population in the language regions and to localize the simulated English speaker: USA v. UK v. Australia.

Length (hidden) In preliminary experiments, we observed that the generated content exceeds the length of typical social media posts. Thus, we explicitly prompt the model to generate content differing in word count. We choose for each generation a random textual modifier to specify the scope: ”few-word”, ”single-sentence”, ”short”, and ”long”. The modifier is a hidden variable and does not influence the annotation or evaluation.

3.2 Data

We generated 1000 equally distributed samples across the above-described dimensions. A native or C2 speaker annotated a subset for their respective language. For the following preliminary analysis, our annotated dataset consists of 600 samples from which we sample a stratified selection, containing 100 per language. The annotation consists of three different dimensions, utilizing a five-point scale from low to high:

1. Topic and 2. Persona Alignment: We generated content for five different political topics and three different political personas. These two dimensions evaluate how closely the sample represents the given topic or the provided persona.

3. Overall Authenticity: This dimension evaluates the overall likeliness of the post to appear based on the topic and persona on the given platform. Thus, the authenticity of the content concerning the provided environment.

The annotation was performed in a graphical interface providing the generated content, including the relevant meta information (topic, persona, platform). See the attached screenshot below.

3.3 Results

Before describing our results in detail, we want to draw attention to two known limitations concerning the expressiveness of our results. Due to the limited number of annotated samples and the lack of multi-annotation (annotator agreement), these results serve as a starting point for more extensive experiments. We suggest possible extensions in the future work section. However, based on the available data, we observe limitations and possibilities confirmed by preceding research.

Model The previously published evaluation of Mistral 7B (Jiang et al., 2023) shows comparable performance to significantly larger models in established benchmark tasks like commonsense reasoning, language understanding, and knowledge-based reasoning. There exists no prevalent dataset for the evaluation of multilingual OSN post-generation. Thus, one research goal of this work is to present a preliminary ranking. Based on three annotation dimensions, our results show Mistral 7B is comparable to GPT-3.5. Generally, both models display a degree of perceived topic alignment, while the persona alignment and overall authenticity yield lower values. We discuss our observations for the persona-related annotation in a separate subsection. With regard to the models, however, we found qualitative

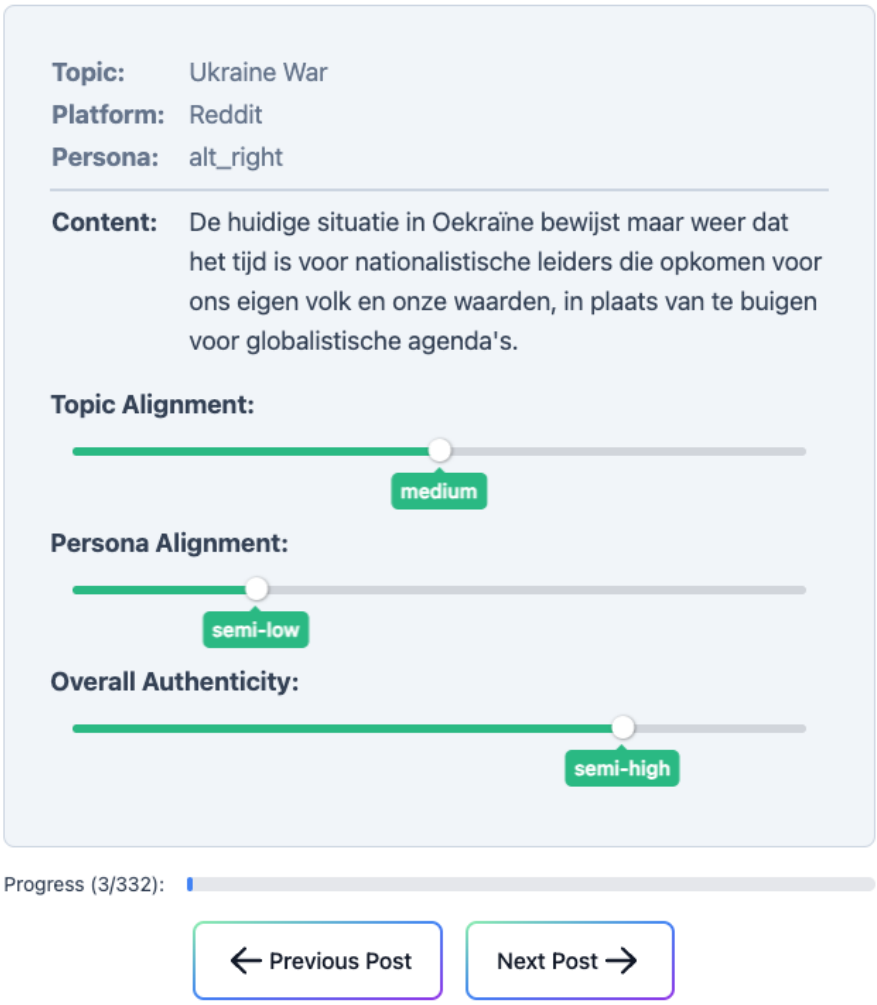


Figure 1: Annotator Interface Example

Metric		GPT3.5-turbo	Mistral-7B-Instruct
Topic	mean	4.788	**4.847**
	std	0.691	0.583
Persona	mean	**4.179**	4.034
	std	1.044	1.118
Authenticity	mean	**3.621**	3.520
	std	1.154	1.256

Table 1: Annotation Results across Models

Persona	Model	Topic	Persona	Authenticity
Alt-Right	GPT-3.5-Turbo	4.660	4.000	3.500
	Mistral-7B-Instruct	**4.860**	**4.060**	**3.720**
Conservative	GPT-3.5-Turbo	4.783	**3.950**	**3.500**
	Mistral-7B-Instruct	**4.804**	3.585	3.170
Liberal	GPT-3.5-Turbo	**4.934**	**4.673**	**3.913**
	Mistral-7B-Instruct	4.867	4.358	3.603

Table 2: Annotation Results across Personas

drawbacks of the generated text. Predominately, the annotators noticed the lack of actuality, like outdated information on COVID-19 or the national political discourse. That is a common theme inre-trained LLMs, as their internal knowledge base is restricted to the data seen during training. In the case of GPT-3.5, the authors used data up to September 2021. Mistral-7B produces English translation for German and Dutch generations. We excluded this phenomenon during the annotation. However, it shows that in contrast to GPT-3.5, it misunderstands our prompt as we do not state the models to provide additional content besides the generated post.

Persona Concerning the perceived persona alignment, we observe a significant difference in the persona and overall authenticity evaluation. The liberal persona shows the highest alignment and generates the most authentic posts. The annotators noted that the alt-right and conservative persona display a more left-leaning worldview. That aligns with previous research (Rozado, 2023) analyzing the political bias of GPT-3.5 and proving that the model has a liberal/left-leaning bias.

Language Modern (western) LLMs are trained on a comprehensive internet crawl. Thus, the training data contains predominately the most-used online language, English. However, these models show great capabilities in translating and generating languages with medium text resources like French, Spanish, or German. Our results underline these findings. We receive the best results when generating En-

Language	Model	Topic	Persona	Authenticity
Dutch	GPT-3.5-turbo	**4.860**	**3.700**	**2.820**
	Mistral-7B-Instruct	4.780	3.400	2.280
German	GPT-3.5-turbo	4.640	4.200	3.900
	Mistral-7B-Instruct	**4.820**	**4.260**	**4.060**
English	GPT-3.5-turbo	4.857	**4.589**	4.089
	Mistral-7B-Instruct	**4.954**	4.500	**4.318**

Table 3: Annotation Results across Languages

glish texts. However, the German generations are nearly on par. For Dutch, a low-resource language, the perceived authenticity is significantly lower. Besides incorrect phrasing and atypical word usage, both models lack built-in knowledge of countries speaking Dutch. The annotators noticed that the generation aligns with the given topic, but the content shows an American-centrism in terms of political viewpoint and consideration of national circumstances.

Implications on TWON The mentioned experiment and the proposed future development impact the research around TWON on multiple levels. First, it allows us to select a suitable candidate model for our simulation phase. As confirmed by our results, we can assume that the generated content feels authentic for users to interact with. However, the results suggest that simulation in low-resource languages – confirmed for Dutch and induced to Serbian – may yield significantly worse generated content. Discussing the issue is necessary as a simulation in Serbian is part of the initial research proposal. In a broader sense, the quality of low-resource languages and the American centrism in the generated content highlight the need for a more EU-centered approach to LLMs. Creating such a model, potentially based on the data collected during our simulations, displays a relevant future work proposal to enhance the quality of EU-based social media analysis. Further, we can reuse the so far generated content as a starting seed for the first simulation. In contrast to a tabula rasa start, a pre-populated network may improve initial user engagement.

Research Insight

The observed performance disparities across languages provided crucial insights for the project's implementation strategy, particularly regarding the challenges of simulating authentic discourse in non-English European contexts. These findings directly informed our subsequent research directions, highlighting the need for language-specific fine-tuning approaches and more sophisticated evaluation metrics. The experiment also established important methodological precedents for assessing model performance across multiple dimensions simultaneously, contributing valuable protocols for ongoing evaluation within the broader TWON framework.

4 Testing LLMs Alignment on Moral Foundations

Research Insight

Having established the technical capabilities of LLMs for generating social media content, our research next addressed the critical question of whether these models could authentically represent diverse ideological perspectives. This investigation into moral foundations alignment was essential for determining the extent to which LLM-based agents could serve as valid proxies for human users with specific political orientations—a fundamental requirement for simulating realistic social network dynamics. Through standardized moral foundation questionnaires, we sought to quantify the degree to which models could consistently maintain ideological coherence when prompted with different personas.

The advancements of Large Language Models (LLMs) not only flooded the consumer market (Teubner et al., 2023) but also academia with text as a research subject (Tiunova and Muñoz, 2023). The abilities of these systems range from classifying and extracting information from unstructured inputs (Xu et al., 2023) to unrestricted text generation adapted to different styles (Bhandarkar et al., 2024). Contemporary research in the social sciences aims to utilize the capabilities to generate content tailored to individual user behavior. A common and predominant approach is to provide an abstract textual description of a political ideology (Argyle et al., 2023). It relies on the model's ability to generalize from abstract ideology description to the appropriate response for generative tasks like social media post generation. However, this research presents no factual evidence or framework to verify how consistently a persona-based (personalized) prompting can resemble individuals with specified ideologies. The underlying assumption in these methods is that LLMs can inherently encode ideological perspec-

tives within their trained parameters.

In contrast to assessing a personalized LLM's ideology, approaches exist to implicitly investigate the political leaning of humans through measuring abstract values and beliefs. Differential psychology utilizes Moral Foundation Theory (MFT) to measure an individual's reliance on five distinct foundational aspects of morality (Graham et al., 2009). Each foundation represents a different set of moral concerns and intuitions that can influence people's attitudes toward various social and political issues. In combination with the self-reported ideology, MFT shows a significant correlation along the five axes and ideologies (Hatemi et al., 2019). In scenarios where LLMs serve as proxies for human users, these artificial agents should demonstrate consistent behavior when responding to written surveys or questionnaires. Hence, transferring the ideas from survey-based assessment onto LLMs may verify the machine's understanding of ideologies.

The deployment of LLMs as substitutes for humans appears particularly convenient for online social networks (OSNs), as researchers can design an environment that is task-specific and centered around text (Argyle et al., 2023). Thus, measuring polarization tendencies on a large scale with a reproducible approach seems possible. In the current landscape, where OSN providers restrict access to data and obstruct researchers from conducting data-driven experiments based on real data (Bruns, 2021), the synthetic approach may pose a promising solution. However, when applying new technologies, especially those driven by market interests, we think an unreflected application poses an imminent danger to the quality and validity of research. We argue that a critical analysis of these models in out-of-domain tasks is fundamental to assessing the validity of high-level applications like the simulation of users with LLMs. Without this critical lens, experiments based on synthetic OSN users yield no insight into how close they can resemble real human interaction.

Research Questions & Contributions Our work aims to provide a foundation for analyzing the impact of persona prompt modifications and their alignment in representing the left-right political spectrum. We see this groundwork as necessary to assess the capabilities of LLMs to generalize from abstract ideologies to practical applications like generating personalized content or reactions. We focus our analysis on the following research questions:

RQ$_1$ How consistently do LLMs perform in their factory settings when surveyed with/without personas by only manipulating them through in-context prompting?

RQ$_2$ How closely align LLMs in their factory settings by only manipulating them through in-context prompting to which human participant groups?

4.1 Background

We aim to connect our work to the existing critique of LLMs, with a focus on their application and the perception of their capabilities in terms of language understanding and ability to communicate. Further, we outline the unreflected application of synthetic users in the social sciences as human replacements and critique the expressiveness of those studies.

4.1.1 Not more than stochastic parrots?

Bender et al. (2021) critiqued that language models only manipulated textual content statistically to generate responses that give the impression of language understanding, like a parrot that listens to a myriad of conversations and anticipates how to react accordingly. Current conversational models are published by commercial facilities, with a business model relying on the illusion of models capable of language understanding and human-like conversation skills (Kanbach et al., 2024). Thus, we have two extreme standpoints towards LLMs: a reductionist perspective that considers these models as next-word prediction machines based on matrix multiplication and an anthropomorphic view that attributes human-like qualities to those systems (Bubeck et al., 2023). While we disagree with a (naive) anthropomorphism and current research questions the language understanding capabilities (Dziri et al., 2024), we argue that when utilizing LLMs as human simulacra (Shanahan, 2024), we must assume human-like qualities to a certain degree. Without this assumption, utilizing LLM agents to model interpersonal communication can only yield a shallow copy, a conversation between parroting entities.

4.1.2 LLMs as synthetic characters

The usage of LLMs as human simulacra (representation) began with the application as non-player characters (NPCs) in a Sims-style game world to simulate interpersonal communication and day-to-day lives (Park et al., 2023). The application of LLMs as synthetic characters has expanded beyond gaming environments into various fields of social science research (Argyle et al., 2023). Those disciplines already started to use these models as a replacement in social studies arguing that conditioning through prompting causes the systems to accurately emulate response distributions from a variety of human subgroups (Argyle et al., 2023). While these applications show promise, they also raise significant methodological and ethical questions. Current research raises concerns about potential biases in the training data leading to misrepresentation of certain groups or viewpoints (Abid et al., 2021; Hutchinson et al., 2020). Without a deeper understanding of the model's representations of ideologies, we risk oversimplifying complex human behaviors and social dynamics. Especially as these approaches (Argyle et al.,

2023) ignore that LLMs lack embodiment in the physical world. This disembodied nature means they lack the grounding in physical reality – expressed by cultural contexts, physical environments, and interpersonal relationships – that shapes human cognition, perception, and decision-making (Hussein, 2012).

4.2 Methods

We repeatedly prompt LLMs to answer an MFT questionnaire with different political persona system prompts to nudge the model toward a left-right ideology. Thus, we obtain a population for each model-persona combination that is the base for our variance and cross-human analysis. The populations contain 50 samples. In total, we obtain $2,400$ artificially filled surveys.

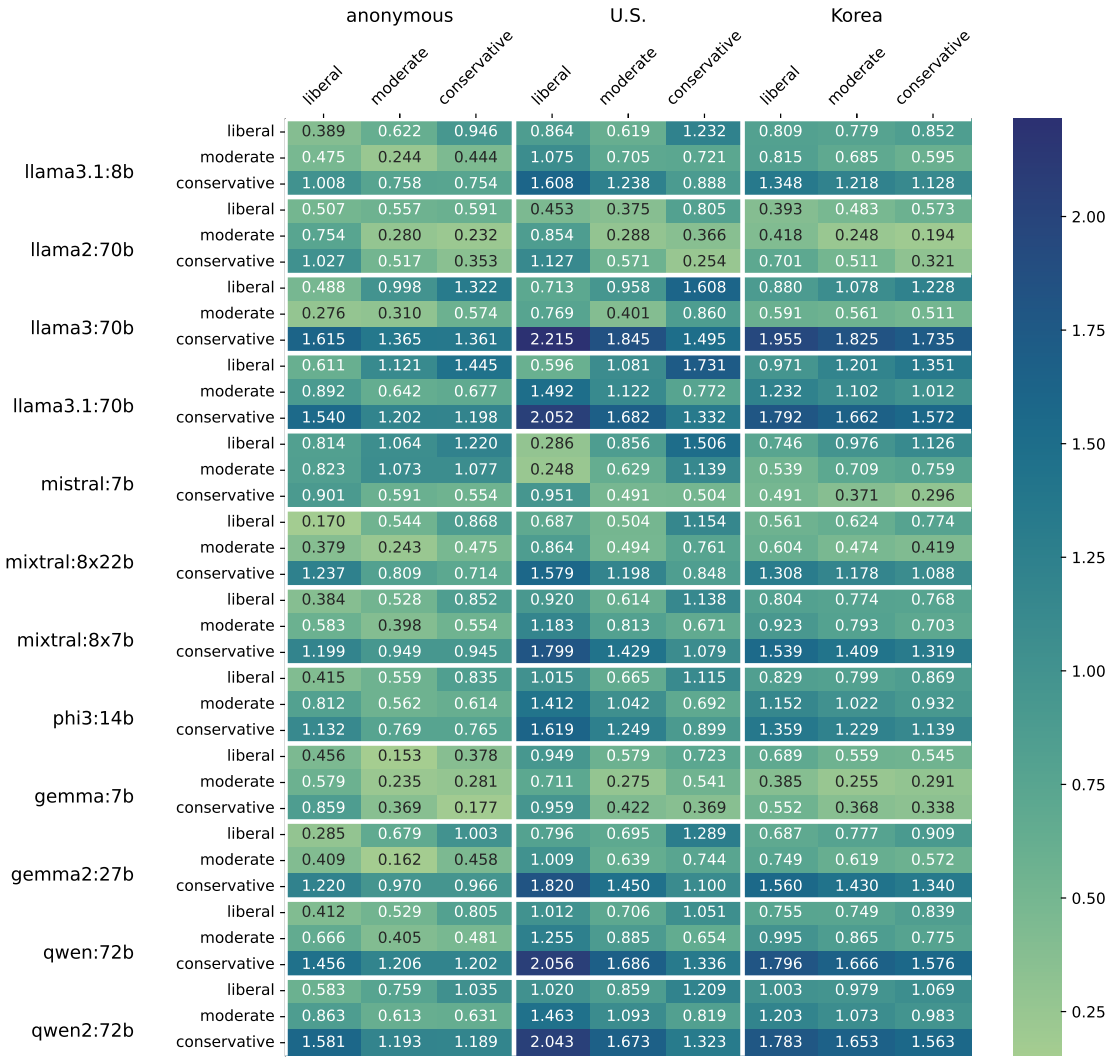
4.2.1 Models

Our research focuses on models with openly available weights that researchers can deploy locally using moderate computational infrastructure — specifically, systems with approximately 80 GB of video memory. These restrictions make our results and experiment pipeline usable for smaller research facilities without access to third-party providers. To broaden the selection across the size of models and their architecture, we include LLMs ranging from 7 B up to 176 B parameters and include models based on a mixture of expert architecture (Du et al., 2022).

4.2.2 Questionnaire

The center of our experiments forms the Moral Foundations Questionnaire (MFQ) originally proposed by (Graham et al., 2009). The MFQ is a psychological assessment tool designed to measure the degree to which individuals rely on five different moral foundations when making moral judgments: care/harm (kindness, gentleness, nurturance), fairness/cheating (justice, rights, autonomy), loyalty/betrayal (solidarity, patriotism, sacrifice), authority/subversion (leadership, fellow-ship, authority), purity/degradation (living in a noble way). The questionnaire consists of 32 items divided into two parts. Moral Relevance: 16 questions asking participants to rate how relevant certain considerations are when making moral judgments. Moral Judgments: 16 questions asking participants to indicate their agreement or disagreement with specific moral statements. Responses are given on a 6-point Likert scale, ranging from 0 to 5. The Moral Relevance scale ranges from ”not at all relevant” to ”extremely relevant”. By using a standardized and well-validated tool like the MFQ, we aim to provide a robust framework for comparing the moral reasoning capabilities of LLMs to those of human participants, while also exploring how

Moral Foundation Questionnaire: Human & LLM Cross-Evaluation



		anonymous			U.S.			Korea		
		liberal	moderate	conservative	liberal	moderate	conservative	liberal	moderate	conservative
llama3.1:8b	liberal	0.389	0.622	0.946	0.864	0.619	1.232	0.809	0.779	0.852
	moderate	0.475	0.244	0.444	1.075	0.705	0.721	0.815	0.685	0.595
	conservative	1.008	0.758	0.754	1.608	1.238	0.888	1.348	1.218	1.128
llama2:70b	liberal	0.507	0.557	0.591	0.453	0.375	0.805	0.393	0.483	0.573
	moderate	0.754	0.280	0.232	0.854	0.288	0.366	0.418	0.248	0.194
	conservative	1.027	0.517	0.353	1.127	0.571	0.254	0.701	0.511	0.321
llama3:70b	liberal	0.488	0.998	1.322	0.713	0.958	1.608	0.880	1.078	1.228
	moderate	0.276	0.310	0.574	0.769	0.401	0.860	0.591	0.561	0.511
	conservative	1.615	1.365	1.361	2.215	1.845	1.495	1.955	1.825	1.735
llama3.1:70b	liberal	0.611	1.121	1.445	0.596	1.081	1.731	0.971	1.201	1.351
	moderate	0.892	0.642	0.677	1.492	1.122	0.772	1.232	1.102	1.012
	conservative	1.540	1.202	1.198	2.052	1.682	1.332	1.792	1.662	1.572
mistral:7b	liberal	0.814	1.064	1.220	0.286	0.856	1.506	0.746	0.976	1.126
	moderate	0.823	1.073	1.077	0.248	0.629	1.139	0.539	0.709	0.759
	conservative	0.901	0.591	0.554	0.951	0.491	0.504	0.491	0.371	0.296
mixtral:8x22b	liberal	0.170	0.544	0.868	0.687	0.504	1.154	0.561	0.624	0.774
	moderate	0.379	0.243	0.475	0.864	0.494	0.761	0.604	0.474	0.419
	conservative	1.237	0.809	0.714	1.579	1.198	0.848	1.308	1.178	1.088
mixtral:8x7b	liberal	0.384	0.528	0.852	0.920	0.614	1.138	0.804	0.774	0.768
	moderate	0.583	0.398	0.554	1.183	0.813	0.671	0.923	0.793	0.703
	conservative	1.199	0.949	0.945	1.799	1.429	1.079	1.539	1.409	1.319
phi3:14b	liberal	0.415	0.559	0.835	1.015	0.665	1.115	0.829	0.799	0.869
	moderate	0.812	0.562	0.614	1.412	1.042	0.692	1.152	1.022	0.932
	conservative	1.132	0.769	0.765	1.619	1.249	0.899	1.359	1.229	1.139
gemma:7b	liberal	0.456	0.153	0.378	0.949	0.579	0.723	0.689	0.559	0.545
	moderate	0.579	0.235	0.281	0.711	0.275	0.541	0.385	0.255	0.291
	conservative	0.859	0.369	0.177	0.959	0.422	0.369	0.552	0.368	0.338
gemma2:27b	liberal	0.285	0.679	1.003	0.796	0.695	1.289	0.687	0.777	0.909
	moderate	0.409	0.162	0.458	1.009	0.639	0.744	0.749	0.619	0.572
	conservative	1.220	0.970	0.966	1.820	1.450	1.100	1.560	1.430	1.340
qwen:72b	liberal	0.412	0.529	0.805	1.012	0.706	1.051	0.755	0.749	0.839
	moderate	0.666	0.405	0.481	1.255	0.885	0.654	0.995	0.865	0.775
	conservative	1.456	1.206	1.202	2.056	1.686	1.336	1.796	1.666	1.576
qwen2:72b	liberal	0.583	0.759	1.035	1.020	0.859	1.209	1.003	0.979	1.069
	moderate	0.863	0.613	0.631	1.463	1.093	0.819	1.203	1.073	0.983
	conservative	1.581	1.193	1.189	2.043	1.673	1.323	1.783	1.653	1.563

Table 4: Absolute difference (lower is better) between the moral foundation scores of the selected models and scores across political ideologies of anonymous participants (Graham et al., 2009) and US-Americans & Koreans (Kim et al., 2012). The scale ranges between 0 (no distance between model and human) and 5 (maximum distance).

Persona Model	Base	Conservative	Liberal	Moderate	MEAN
Gemma:7b	0.073	0.134	0.061	0.057	0.081
Llama2:70b	0.309	0.514	0.422	0.447	0.423
Llama3:70b	0.116	0.062	0.089	0.300	0.141
Mistral:7b	0.259	0.665	0.204	0.489	0.404
Mixtral:8x22b	0.162	0.134	0.112	0.180	0.147
Mixtral:8x7b	0.025	0.037	0.047	0.012	0.030
Qwen:72b	0.108	0.116	0.356	1.122	0.425
MEAN	0.150	0.237	0.184	0.372	0.236

Table 5: Response variance aggregated across questions by model and persona.

different prompting strategies might influence these capabilities.

4.2.3 Prompting

The intention of our work is to assess synthetic surveys and evaluate the alignment between participants and language models. Thus, we opt for a simple prompt, containing only the task and an optional persona stating the political and ethical ideology. With the reduction to the keywords of the political ideology, we force the system to tap into its built-in concepts without modifying them in context (Wei et al., 2021). The persona description prompts the model to represent the opinion of a ”politically and ethically {Conservative | Moderate | Liberal}”. We prompt the model on each question individually paired with the task description.

4.3 Results

Our response variance results (Table 5) show a significant difference between the different models and personas. While Mistral 8x7b shows the highest stability with the lowest variance (0.030), Qwen 72b has a 14 times higher (0.425) variance. Also, adding ideological personas consistently increased response variance, with moderate personas (0.372) showing the most significant deviation from baseline responses (0.150). Our cross-evaluation (Table 4) shows that on average the models exhibited left-leaning bias as reported for the GPT-family (McGee, 2023; Rutinowski et al., 2024).

4.4 Discussion

The inconsistency in model responses, particularly evident in Qwen, raises concerns about the reliability of using LLMs as proxies for human participants in social science research. Also, larger models did not consistently outperform smaller ones in our study. This suggests that model size alone does not

guarantee better performance in tasks requiring a nuanced understanding of human values and beliefs. While our results show that Mixtral produces the most human-like and consistent responses across our model selection, the overall alignment between model outputs and human participant ideologies is limited. It highlights the restriction of prompting approaches to align LLMs with complex human belief systems and indicates that these systems do not have a built-in concept of those ideologies, at least not capturable using our proposed approach.

Political Biases ChatGPT reported as left-leaning and progressive(McGee, 2023; Rutinowski et al., 2024), aligns with the findings that our models show a smaller distance to liberal groups than conservative (Table 4). The left-leaning tendencies could lead to over-representing progressive viewpoints and potentially skewing the discourse. This skewing may not be limited to generated content but also to the general interaction pattern, where the agents might engage more frequently with liberal content, boosting its visibility and perceived popularity.

RQ$_1$ LLMs showed varying levels of consistency in their performance when surveyed with and without personas. These findings suggest that LLMs' consistency can be significantly affected by incorporating textual personas, and this effect varies considerably across different models. The observed variation could be due to biases in training data, limitations in model architecture, or fundamental challenges in representing complex moral concepts computationally.

RQ$_2$ While Mixtral models showed the best overall alignment, there is no clear, consistent pattern of specific model-persona combinations aligning well with particular human participant groups. This suggests that simple prompt-based persona modifications may not be sufficient to accurately represent diverse human ideologies and moral foundations. The observed misalignment between model outputs and human responses may be partially attributed to anthropic bias in LLMs. These models, trained on human-generated data, may inadvertently reflect and amplify certain human biases or cultural assumptions, limiting their ability to accurately represent diverse moral and political perspectives. Based on our observations, we conclude that we can hardly speak by imitating human ideologies with language models to generate text. The criticized work on human simulacra (Argyle et al., 2023; Park et al., 2023) merely investigates the generated content or opinions on a superficial level but omits a questioning of the validity of LLMs' representation of thought processes and accuracy in reproducing ideologies.

Variance: The lower the better? The preceding results and discussion focus on the observed variance in the collected data. Our analysis assumes a lower variance as the favorable outcome. We assume

that a lower variance indicates a more robust alignment with the given ideology and, thus, a more reliable behavior when answering the questionnaire. However, when assuming LLMs as human simulacra, this reliability may not be favorable, and we may also use the variance as a comparison metric to the uncertainty in human behavior.

4.5 Conclusion

Our results indicate that researchers must remain cautious and critical when applying these models in social science contexts, considering ethical implications and potential limitations. Based on our findings, we argue that utilizing persona-modified LLMs as human simulacra cannot represent abstract political ideologies and thus yield an inadequate representation of human discourses that merely simulate a superficial discourse. Reducing interpersonal communication to worldly disembodied chatbots in a black-box scenario seems like a dangerous method of riding the AI hype train.

Research Insight

The observed inconsistencies in model responses revealed important limitations in using simple prompting techniques to simulate politically diverse user populations. These findings significantly informed the subsequent development of our data-driven alignment framework, highlighting the need for more sophisticated approaches to persona modeling. The moral foundations assessment thus contributed crucial boundary conditions for the project's simulation objectives and established important methodological considerations for interpreting results from LLM-generated social interactions in political contexts.

5 Towards data-driven OSN-user alignment

Research Insight

Drawing upon insights from our previous experiments, this phase of research focused on developing a comprehensive formal framework for creating empirically realistic simulations of online social networks. This methodological contribution represents the culmination of our technical investigations, providing a structured approach to quantifying and optimizing the realism of simulated user behaviors. By separating different interaction modalities and implementing specialized machine learning methods for each, the framework addresses the complex, multi-faceted nature of social media communication.

Across the world, there are ongoing legislative initiatives to regulate social media platforms. For example, Australia and Spain plan to restrict the use of social networks by minors under the age of 16[1], and TikTok is already banned in several countries due to the spread of propaganda, misinformation, and harmful content.According to the Digital Service Act (DSA) of the EU, providers of very large online platforms need to ”identify, analyze and assess any systemic risks [...] from the use made of their services” like ”any actual or foreseeable negative effects on civic discourse [...].”

However, obtaining realistic quantitative evidence concerning the risks that online platforms pose, and specifically what role platform mechanisms, like ranking and recommendation play, is still being debated in computational social and communication science. The key challenge to obtaining robust insights is to produce counterfactual evidence that tests whether alternative network mechanisms would have had a different (more favorable) outcome. This can best be achieved by simulation, since alternative designs can be tested without the involvement of real users.

However, the validity of the insights obtained by simulation can only be as good as it represents reality. With the advent of Large Language Models (LLMs) for the imitation of persons in social simulation, the complexity of communication in social simulation has become arguably more realistic compared to previously used symbolic agent communication languages. Whether LLM-based simulation can replicate social science studies with human participants to a degree sufficient for robust scientific insights needs an experimental design that improves reproducibility and provides a measurement of empirical realism. Testing under which conditions and to what extent agents can mimic human behavior is crucial to validate simulation-based empirical findings.

This paper contributes to the rigor of conducting simulations of social networks with LLMs by:

- Formalizing social networks to i) quantify and standardize their simulation and ii) allow to benchmark the empirical realism of the simulation.
- Providing different approaches and benchmark data sets for the imitation of different types of user communication on $\mathbb{X}$, like posting and replying.
- Evaluating the empirical realism and identifying key findings about the potential and limitations of mimicking users with the help of LLMs.

5.1 Preliminaries

There is a long history of distributed artificial intelligence research and multi-agent systems in modeling agents to resemble human decision-making processes Weiss (1999). Those traditional approaches

[1]Australia approves social media ban on under-16s: `https://www.bbc.com/news/articles/c89vjj0lxx9o`

define top-down cognitive models of white-box agents with abstract concepts and deductive reasoning processes that control the agent's behavior. Similarly, the rational choice model and game theoretic models are concerned with explicit motives and perceived restrictions underlying human behavior. The social sciences have come up with many alternative explanations of human behavior. However, all of those theories tend to be rather inaccurate when it comes to predicting human behavior. In contrast, this paper is solely concerned with mimicking the communication of users as precisely as possible and does not care to explain or control the behavior. This implies that the agents are generating natural language and not symbolic agent languages such as FIPA ACL O'Brien and Nicol (1998).

The usage of LLMs as human simulacra (representation) began with the application as non-player characters (NPCs) in a Sims-style[2] game world to simulate interpersonal communication and day-to-day lives Park et al. (2023). The results showed superficially authentic and believable human behavior. Current research interest revolves around improving those agents in a technical sense, by refining prompt schemes and model-internal feedback loops Wang et al. (2024). However, the application of LLMs as synthetic characters has expanded beyond gaming environments to various fields of social science research Argyle et al. (2023). Researchers are increasingly exploring the potential of these models to simulate human participants in studies, particularly in contexts where the obtaining of real-world data is challenging or ethically complex. Those disciplines have already started to use these models as replacements in social studies arguing that conditioning through prompting causes the systems to accurately emulate response distributions from a variety of human subgroups Argyle et al. (2023).

Although these applications show promise, they also raise significant methodological and ethical questions. The reliability and validity of using LLMs to represent human behavior and cognition is still subject to debate. Current research raises concerns about potential biases in training data that lead to misrepresentation of certain groups or points of view Abid et al. (2021); Hutchinson et al. (2020). Furthermore, the use of LLMs as replacements for human participants in social research raises ethical considerations about informed consent and the potential for misuse or misinterpretation of results. Without a deeper understanding of the model's representations of ideologies, we risk oversimplifying complex human behaviors and social dynamics. Especially since these approaches ignore the LLMs lack of embodiment in the physical world Argyle et al. (2023). This disembodied nature means they lack the grounding in physical reality – expressed by cultural contexts, physical environments, and interpersonal relationships – that shapes human cognition, perception, and decision making Hussein (2012).

Although less complex than modeling communicative behavior, research on LLMs generating synthetic public opinion demonstrates whether base LLMs can accurately represent opinions in a socio-

[2] The Sims is a series of life simulation video games developed by Maxis and published by Electronic Arts

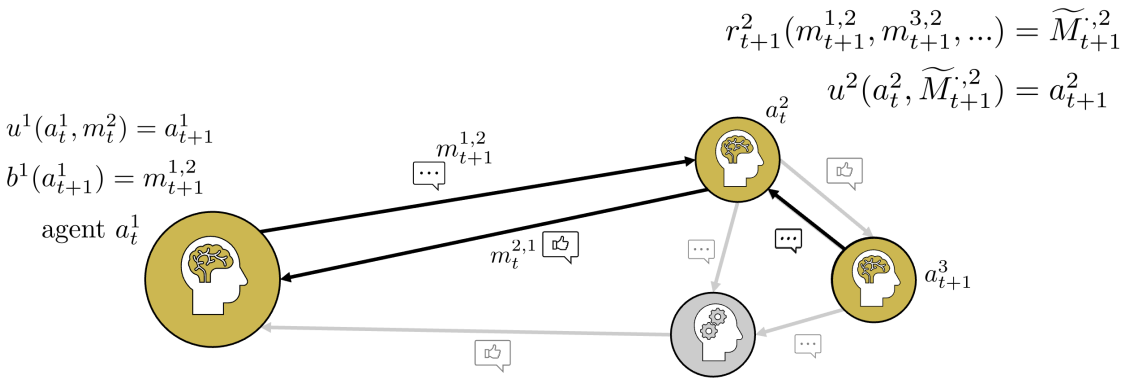


Figure 2: Illustrating example: Agent 2 sends a message to agent 1, which updates its model and generates a reply. Agent 2 receives a curated list of all incoming messages in $t+1$ and updates its model based on that.

cultural context. Although LLMs show an excessive ability to replicate US personas - closely approximating real world opinion distributions - they underperform when being tasked with replicating more complex non-English populations Argyle et al. (2023); Ma et al. (2024). Ma et al. demonstrate sensitivity to variable inclusion/exclusion. Similarly Tjuatja et al. (2024) report significant behavioral changes in LLM trained with RLHF due to question perturbations, attributing this behavior to the training scheme, stressing the need to adjust LLMs to case-specific data to improve grounding in reality.

5.2 Formalization

In this section, we provide a concise formal framework for modeling twins of online social networks. Both the users (Sect. 5.2.1) and the network mechanics (Sect. 5.2.2) are captured. Fig. 2 provides an illustrating example.

5.2.1 Modeling Agents

Let A be a set of agents, i.e., social media users, where a^i_t is the state of the i-th agent at time t.[3]

M is a set of messages, where $m^{(i,j)}_t$ is a message from agent i to agent j at time t. Consequently, $m^{(i,\cdot)}_t$ indicates a message that agent i transmits to all other agents and $M^{(\cdot,i)}_t$ indicates all messages sent to agent i.[4] Accordingly, M^i_t indicates all messages that agent i sends and receives at time t.

The state a^i_t of the i-th agent at time t is defined by the discourse it has been involved in up to this point: $(M^i_0, M^i_1, ..., M^i_t)$. The agent's state is updated according to the function $u^i(a^i_t, M^i_t) = a^i_{t+1}$ based on its previous state and the messages sent and received since the last point in time.

[3] The internal state of an agent can be represented by symbolic logic, machine learned functions, cognitive architecture or any other agent formalism that is capable of perceiving its environment and acting upon that environment.

[4] Note, that messages can be of any modality or type (including images, videos, likes, shares,...)

Behavior b	**Agent** A	**Task Type**	**Input** $\widetilde{M}_t^{(\cdot,i)}$	**Output** $m_{t+1}^{(i,\cdot)}$	**Loss** L_b	**Metric(s)** q
Posting	Creator	Generative	Topic	Free Text	Cross Entropy	BLEU, ngram,...
Replying	Reactor	Generative	Post	Free Text	Cross Entropy	BLEU, ngram,...
Replying Likelihood	Reactor	Probability	Post	Interval [0,1]	Binary Cross Entropy	F1, Acc.
(Dis-)Liking Likelihood	Reactor	Similarity	Post	Interval [-1,1]	*untrained*	Cosine Similarity

Table 6: Communicative behaviors modeled, imitated and benchmarked in our empirical study with respect to their empirical realism, as definied in our formal framework.

The communicative behavior $b^i()$ of the agent i is a function that maps its current state a_t^i to the messages $m_{t+1}^{(i,j)}$ he is sending next to agents j: $b_t^i(a_t^i) = M_t^{(i,j)}$.

5.2.2 Modeling Network Mechanics

The mechanics of the communication channel (here, an online social network) is defined by a function $r_t^i()$ that adapts the agent i' to the incoming messages $M_t^{(\cdot,i)}$ to $\widetilde{M}_t^{(\cdot,i)}$, returning a manipulated list of messages. Standard manipulations of the list are filtering or ranking, but could also include the manipulation of content[5] or the addition of messages to the list that have not been directly sent between agents[6]. Note that this function can be personalized and can be adapted over time.

In general, the messages agent i receives at time t are given by:

$$r_t^i(b_t^j(a_t^j)) = \widetilde{M}_t^{(\cdot,i)} \quad \forall j \in A_t^j$$

which in turn triggers agent is response

$$b_{t+1}^i(u^i(a_t^i, \widetilde{M}_t^{(\cdot,i)})) = M_{t+1}^{(i,\cdot)}$$

5.3 Simulating Online Social Networks

Based on the formal framework introduced in Sect. 5.2 we define the required tasks to construct a twin of a real-world online social network by replicating its behavior at the user (Sect. 5.3.1) and system level (Sect. 5.3.1). Finally, this allows us to check the empirical realism of such a TWON (Sect. 5.3.3).

5.3.1 Imitating User Behavior

The task of machine learning an agent-based simulation of human social media communication behavior at user level is to estimate the function b^i of each agent i, which given real-world observations

[5] e.g., by adding community notes or fake news warnings
[6] e.g., advertisements, paid content, trending content,...

of discourses $(M_0^i, M_1^i, ..., M_t^i)$ predicts its next message m_{t+1}^i. Thus, the objective can be formulated as minimizing a loss function

$$\min L_b(\widehat{m}_{t+1}^i, m_{t+1}^i) \mapsto \mathbb{R}$$

that compares the predicted $\widehat{m}_{t+1}^i$ to its observed value.[7]

5.3.2 Replicating Network Mechanics

The task of replicating the system-level mechanics of a social network from observations is to estimate the function r^i of each agent i, given its observed messages $(\widetilde{M}_1^i, \widetilde{M}_1^i, ..., \widetilde{M}_t^i)$.

This objective can be formulated as minimizing the loss

$$\min L_r(\widehat{M}_t^{(\cdot,i)}, \widetilde{M}_t^{(\cdot,i)}) \mapsto \mathbb{R}$$

that compares the predicted $\widehat{M}_t^{(\cdot,i)}$ with its observed value.[8]

5.3.3 Simulating Social Networks

Finally, a simulation task for investigating user communication on social networks can be stated as follows: Given (an estimation of) b and r generate a discourse $M_{t+1}, ..., M_{t+n}$ and evaluate it according to a discourse metric q, for example, the degree of outrage or hate speech, mapping

$$q(M_{t+1}, ..., M_{t+n}) \mapsto \mathbb{R}$$

Any findings obtained by q can be put into perspective by L_b and L_r as they quantify the empirical realism of the simulation of a social network and consequently qualify the validity of q. L_b and L_r can be interpreted as a confidence score for any empirical findings obtained by simulating social networks and should be reported together with empirical findings obtained by simulation.

5.4 Machine-learned/Approximated User Actions

In the following sections, we instantiate the formal framework described in the previous sections based on observations from the social network $\mathbb{X}$. Our focus is on estimating a set of users A of $\mathbb{X}$, specifically

[7] Any text comparison metric, like BLEU, cross-entropy, but also higher level metrics could be applied here. In addition, other metrics can be used to measure the predictive accuracy of discrete actions such as "liking".

[8] Here, a ranking metric like MRR or NDCG seems best suited.

their communicative behavior b. We do so by minimizing $L_b(\widehat{m}_{t+1})$, where m can for example be a message of type *post*, *reply*, or *like*. This work is *not* concerned with the estimation of network mechanics and the distribution and targeting of messages modeled by r.

The imitation of agent behavior b is further separated into distinct subtasks: *Posting* behavior is characteristic for the content creator user group and the *Replying* behavior for reactive users. In our data, the former category includes politicians, policy makers, and institutional news outlets, characterized by their role in initiating discussions and setting conversational agendas. The latter comprises individual users whose primary mode of engagement involves responding to and amplifying existing content through platform-specific mechanics such as replies, retweets, and reactions. This asymmetric structure reflects the empirically observed power-law distribution of participation in social networks, where a small percentage of users generate most of the original content (Carron-Arthur et al., 2014; Sun et al., 2014).

An overview of the different agent behaviors that we empirically evaluate in the next sections is provided in Table 6 and is described below.

5.4.1 Imitating Posting Behavior

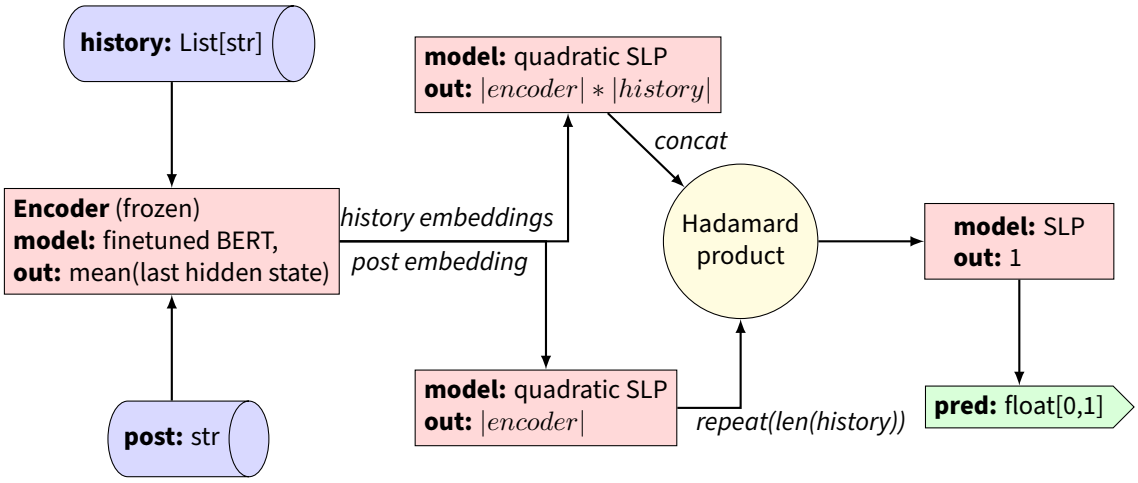
We model party-specific content generation patterns through the extracted topics and the personal information of the individuals (name and party affiliation). Given a specified topic, the model learns to generate content that reflects both the broad party stance and individual variations in expression style.

5.4.2 Imitating Replying Behavior

The reply generation system operates on a context-aware framework that processes historical interaction patterns. The historical interaction context is represented as a sequence of <Post, Reply> pairs, encoded as few shot examples during the alignment process. The model maintains consistent user behavior by implementing a constraint mechanism that prevents top-level post-generation, restricting outputs to reply-only contexts.

5.4.3 Estimating Replying Likelihood

The goal of this task is to predict whether a synthetic social network user would respond to a post based on their history of interaction. The system processes historical interactions and a current post using a fine-tuned BERT model to generate embeddings by calculating the pooled mean of its input representations. Historical interactions and the post are separately processed through quadratic linear layers



SLP: single layer perceptron followed ReLU activation

Figure 3: Workflow for approximating the reply-action likelihood with sentence embeddings and cosine similarity

with ReLU activation, maintaining dimensions related to historical data and encoder size. The post embedding is repeated to match each historical entry and combined using element-wise multiplication with the history representation, establishing an interaction between past patterns and current post features. These combined representations are fed into a classification layer, producing a prediction score between 0 and 1 to indicate the likelihood of a synthetic user replying to the post.

5.5 Estimating (Dis-)Liking Likelihood

The system evaluates a new post by comparing it with the user's historical interactions. It uses a fine-tuned BERT model to generate vector embeddings for both the historical data and the new post, capturing their semantic meanings. A centroid is calculated from the user's past interactions to represent their average interests. The system then computes cosine similarity between this centroid and the new post's embedding, producing a prediction score from -1 to 1, indicating the user's potential preference for the new content. In addition, we define a threshold value $t = 0.5$ and the function $\mathcal{L}$ that takes the output $\hat{y}$ of our model defined above and returns the corresponding action. As our data does not contain information on this specific action, we have no described experiments to test the performance and validity of this approach.

$$\mathcal{L}(\hat{y}) = \left\{ \begin{array}{ll} \text{"}dislike\text{"}, & \text{if } \hat{y} \leq -t \\ \text{"}ignore\text{"}, & \text{if } \hat{y} \in (-t, t) \\ \text{"}like\text{"}, & \text{if } \hat{y} \geq t \end{array} \right\}$$

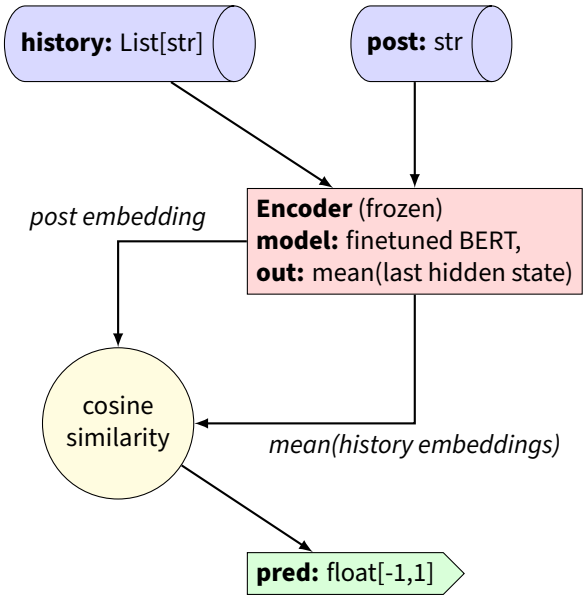


Figure 4: Workflow for approximating the like-action likelihood with sentence embeddings and cosine similarity

5.6 Experiments

In this Section, we describe the data set used to train and evaluate our approaches for imitating the types of user behavior described in the previous Section. Furthermore, the machine learning processes for fitting the models to the data is described.

Dataset Our experiments are based on two datasets – English and German – collected from $\mathbb{X}$, previously known as Twitter. The sets are collected around keywords concerning the political discourses in the US and Germany during the first half of 2023. The samples contain two types of content: a) Tweets (posts) from delegates of the national parliament concerning political decisions (DE: 155.000 | EN: 930.000), and b) replies from regular Twitter users towards these decisions or opinions (DE: 185.000 | EN: $17.800.000$). In addition, we labeled the topics of posts and replies using zero-shot prompt classification with Llama-3.1:70b. This yields two distinct agent types, the a) political actor for posts and the b) unspecified Twitter user for replies, both across two languages.

Preprocessing We apply moderate content filtering to enhance the quality of the selected data:

External Source We remove samples that include URLs to external sources. Although this significantly reduces the number of samples, our models cannot access this information during the training and inference process.

Retweets Tweets starting with *"RT"* mark a shared post. The author does not add new content, but replicates the original text and shares it with their followers. We exclude these samples because we cannot verify if the shared content is in line with the retweeting author's opinion.

Content Length We aim to focus on content that contains arguments or opinions, particularly given the later applied discourse metrics. Thus, we remove samples with less than 32 characters based on our heuristic assumption that a specified length is necessary to convey an opinion.

After the initial pre-processing, we further reduce both datasets to retrieve only the most active users measured by the number of posts and replies. This is required for modeling both tasks on a user-based level by providing user examples (few-shot) during prompting.

Content Generation Alignment Pipeline We train two model adapters on top of Llama-3.2-3B-Instruct (Dubey et al., 2024). Both models are trained using the supervised fine-tuning (SFT) paradigm through the transformer reinforcement learning (TRL) library using the default pipeline. The trained adapters are unmodified LoRA (Hu et al., 2021) matrices provide by the parameter-efficient fine-tuning (PEFT) package. We optimize for two objectives (posting and replying) with contents from the original users as the labeled data.

Behavior Likelihood Alignment Pipeline For the decision task, we train our dual input model using embeddings generated by a frozen BERT-based encoder that was pretrained to capture semantic representations of tweets (Zhang et al., 2023). We optimize with AdamW (Loshchilov et al., 2017) using the same data set as for the generative tasks, making the assumption that when a user has not commented on tweets in our collection, this represents an active decision not to reply. To maintain balance, we sample an equal number of negative and positive instances for both languages.

Data and Code Availability The complete technical pipeline for this study is publicly available on GitHub (anonymized during review) at `https://anonymous.4open.science/r/TWON-Agents-D0E2/`. This repository includes the source code for all the data processing, training, and evaluation steps described in this paper. The raw and processed datasets used in this study will be made available after acceptance.

5.7 Results

We evaluated the approaches for the different imitation tasks as described in Sect. 5.6 and Sect. 5.4 to assess their empirical realism.

Generative Task Results

	English			
	Post (persona-based)		**Reply (history-based)**	
	in-context	fine-tuned	in-context	fine-tuned
BLEU [a]	0.015 (± 0.002)	**0.085** (± 0.017)	0.019 (± 0.005)	**0.734** (± 0.047)
unigram [b]	0.157 (± 0.005)	**0.294** (± 0.016)	0.167 (± 0.008)	**0.782** (± 0.046)
bigram [b]	0.022 (± 0.002)	**0.095** (± 0.015)	0.021 (± 0.004)	**0.731** (± 0.053)
length ratio [c]	1.632 (± 0.053)	**1.001** (± 0.054)	1.324 (± 0.058)	**1.003** (± 0.019)
TweetEval [d]				
topics	0.587 (± 0.189)	0.645 (± 0.207)	0.337 (± 0.216)	**0.789** (± 0.193)
emotions	0.437 (± 0.108)	**0.562** (± 0.105)	0.336 (± 0.117)	**0.809** (± 0.076)
sentiment	0.435 (± 0.073)	**0.619** (± 0.083)	0.386 (± 0.113)	**0.841** (± 0.057)
offensive	0.491 (± 0.061)	**0.563** (± 0.050)	0.379 (± 0.098)	**0.785** (± 0.058)
hate	0.327 (± 0.178)	0.200 (± 0.085)	0.231 (± 0.191)	**0.625** (± 0.216)
irony	0.190 (± 0.085)	**0.353** (± 0.085)	0.169 (± 0.120)	**0.737** (± 0.083)
embed. dist. [e]	2.018 (± 0.101)	**1.461** (± 0.085)	2.006 (± 0.162)	**0.544** (± 0.117)
	German			
BLEU [a]	0.009 (± 0.002)	**0.025** (± 0.007)	0.005 (± 0.001)	0.021 (± 0.031)
unigram [b]	0.176 (± 0.010)	**0.238** (± 0.016)	0.099 (± 0.006)	0.148 (± 0.050)
bigram [b]	0.016 (± 0.001)	**0.035** (± 0.003)	0.006 (± 0.003)	0.032 (± 0.044)
length ratio [c]	1.142 (± 0.044)	**1.023** (± 0.049)	1.961 (± 0.234)	**0.785** (± 0.117)
TweetEval [d]				
topics	0.238 (± 0.300)	0.340 (± 0.301)	0.056 (± 0.225)	0.039 (± 0.225)
emotions	0.123 (± 0.278)	0.220 (± 0.251)	0.060 (± 0.215)	0.125 (± 0.261)
sentiment	0.079 (± 0.276)	0.255 (± 0.200)	0.066 (± 0.196)	0.246 (± 0.264)
offensive	0.104 (± 0.215)	0.095 (± 0.239)	0.141 (± 0.200)	0.278 (± 0.350)
hate	0.234 (± 0.340)	0.477 (± 0.247)	0.163 (± 0.212)	0.044 (± 0.160)
irony	-0.084 (± 0.242)	0.118 (± 0.153)	0.166 (± 0.226)	0.074 (± 0.163)
embed. dist. [e]	2.530 (± 0.460)	**1.499** (± 0.195)	3.855 (± 0.437)	**2.891** (± 0.351)

[a] BLEU with smoothing (Lin and Och, 2004), [b] average precision score, [c] token ratio between generated and original content, [d] TweetEval (Barbieri et al., 2020) classifications evaluated with pearson correlation coefficient (higher is better) and aggregated across subclasses, [e] pairwise embedding distance (lower is better) based on TwHIN-BERT (Zhang et al., 2023) [CLS] token

Table 7: Comparison of the base model - Llama-3.2-3B (Dubey et al., 2024) - (in-context prompting) and the aligned version - adapter fine-tuning (Yu et al., 2023) - on our German and English Twitter politician and follower datasets for the posting and replying task evaluated on $n = 100$ independent random samples for $k = 10$ repetitions not seen during training. Bold marks the in-class (language, task) best values if significant by standard deviation.

5.8 Evaluation of Imitated Posting and Replying Behavior

The text-generation-based tasks are evaluated with two different approaches, in two different languages, using three text comparison metrics (see. Table 7).

In-context vs Fine-tuning An initial observation is that BLEU scores in all in-context settings are low. This is also true for the n-gram metrics. However, the fine-tuning-based approach shows significant improvements specifically in the English reply task, but also for unigram matches and the overall length ratio. Beyond such token-overlap-metrics, we tested semantic similarity metrics: we embedded all samples using Twitter-fine-tuned BERT (Zhang et al., 2023) and calculated the distance between generations and original samples. Again, the results generated from the fine-tuned model show a significantly shorter distance to its original counterpart compared to the in-context model. The correlation between the predicted emotions in the original and generated samples using a BERT-based multilabel emotion classifier (Barbieri et al., 2020) is less clear. The fine-tuned approach only marginally outperforms the prompt-only approach on the posting task. However, the English fine-tuned reply generation task showed by far the best performance with significant improvements in all metrics. The BLEU score increased substantially from 0.019 to 0.734, unigram precision improved from 0.167 to 0.782, and the embedding distance was reduced significantly from 2.006 to 0.544. The model achieved a near-perfect length ratio of 1.003 and demonstrated strong TweetEval correlations across all categories.

English vs German In general, we see the same improvements through fine-tuning compared to in-context with respect to the German data in the post-generation task: The BLEU score increases and unigram precision improves. Also, the embedding distance was reduced from 2.530 to 1.499, and the length ratio moved closer to 1.0 after fine-tuning (1.023 vs 1.142). However, compared to English, the TweetEval metrics proved generally unreliable for German content, showing inconsistent patterns with high standard deviations. This suggests fundamental limitations in applying English-trained evaluation metrics to German content. Similarly, the German reply generation (history-based) task shows low empirical realism on all metrics compared to English and regardless of fine-tuning. There were minimal improvements in BLEU scores from 0.005 to 0.021, with high variance in metrics, exemplified by the bigram standard deviation of 0.044. The embedding distance remains high even after fine-tuning at 2.891.

Language	Class	Precision	Recall	F1-Score	Support
German	Ignored	0.741	0.630	0.681	500
	Replied	0.678	0.780	0.726	500
	Weighted Avg	**0.710**	**0.705**	**0.703**	1000
English	Ignored	0.978	0.978	0.978	500
	Replied	0.978	0.978	0.978	500
	Weighted Avg	**0.978**	**0.978**	**0.978**	1000

Table 8: Replying Likelihood metrics for German and English data. TwHIN-BERT (Zhang et al., 2023) (frozen) utilized to embed the individual samples. The train/test combinations are chosen by the iteration with the best test F1-Score.

5.9 Evaluation of Replying Likelihood Behavior

Similar to our evaluation of the generative tasks, we observe substantial performance disparities between English and German models (Table 8). Even on the German training set, our model achieves only moderate alignment with an F1-Score of 0.666, in stark contrast to the English experiment, where the prediction scores are almost perfect.This performance gap suggests that the BERT encoder employed in our study represents English samples with semantically richer embeddings than their German counterparts.

5.10 Key Findings

Several key findings emerge from this analysis. First, language-specific performance shows that English models significantly outperform German models across all metrics, with TweetEval metrics proving more reliable for English content while German models show limited success, particularly in reply generation. Second, task-specific patterns reveal that fine-tuning shows consistent improvements for post-generation in both languages. However, reply generation varies dramatically between languages, with English reply generation achieving remarkably high scores. Third, the impact of data volume is evident, with better performance in English likely due to larger LLM pre-training datasets Fourth, in terms of evaluation metrics, BLEU scores and embedding distances provide consistent signals, TweetEval metrics are only meaningful for English content, and the length ratio serves as an intuitive indicator of style authenticity.

5.11 Imitation Limitations

The stark contrast between English and German model performance, particularly in reply generation tasks, underscores the importance of data volume and language-specific considerations when utilizing generative-agent-based modeling for imitating user behavior on social networks. The success of

the English models demonstrates the potential of the approach, while the limitations faced by the German models highlight that they require additional training and optimizations to provide robust levels of empirical realism.

Our experimental setup has further limitations regarding both data selection and the system's capacity to generate discourses containing well-structured standpoints and arguments. Although our models demonstrate such capabilities under certain conditions, their performance characteristics primarily reflect the behavioral patterns of the most active users within our dataset, thus mirroring the distinctive communication dynamics observed in the selected Twitter community. This sampling bias raises questions about the generalizability of our findings in different contexts of social networks and user populations.

Research Insight

The remarkable performance improvements achieved through fine-tuning, particularly for English content, demonstrate the significant potential of our approach for creating realistic simulations despite the identified challenges. These results establish an important proof-of-concept for the TWON project's core objectives and provide clear technical pathways for future development. The framework also contributes valuable metrics and benchmarks for ongoing evaluation of simulation quality, ensuring that insights derived from these artificial environments maintain scientific validity and policy relevance.

6 Conclusions and Future Work

In this paper we questioned the empirical realism of generative-agent-based modeling for imitating user behavior on social networks. First, we provided a formal framework for building realistic Twins of Online Social Networks (TWONs). Second, we instantiated this framework with the purpose of mimicking user behavior based on data from 𝕏 in English and German. Third, we benchmarked the empirical realism of agents, imitating actual users.

6.1 Key Recommendations

Our empirical results provide several key recommendations for conducting social simulations based on generative-agent-based modeling:

First, simulation models should be validated with respect to their empirical realism before conducting simulations. Results of a simulation should always be put into perspective to the empirical realism

of all components in the simulation.

Second, simulations should be performed in the same setting in which the simulation components were fitted and validated. The stark difference between English and German performance demonstrates this. The English models performed significantly closer to their real-world counterparts and produced more stable results. Changing the setting without retraining and validation can lead to unreliable outcomes of a simulation.

Third, fine-tuning of the simulation components is required to obtain sufficient levels of empirical realism. In-context prompting of LLMs, as is done in most of the related work, specifically in areas like psychology, social sciences, or media studies is often not sufficient to guarantee a simulated behavior that is close to reality.

6.2 Future Research

This paper is a first step towards more robust empirical research designs and protocols for studying real-world social networks by simulating users with generative-agent-based models. Although we established general formal models and initial empirical insights, further research is required. The heterogeneous nature of social networks suggests that discourse patterns can vary significantly between different communities, each with its own linguistic norms, interaction styles, and argumentation preferences. This diversity presents challenges both methodologically and theoretically. Methodologically, there is uncertainty about whether LLMs are adequate to capture these variations or if more specialized models tailored to specific communities are needed. Theoretically, this leads to a broader inquiry about whether social media discourse's inherent heterogeneity necessitates a community-specific modeling approach rather than universal models.

6.3 Limitations

We have already extensively discussed the limitations of our framework and experiments throughout this paper, specifically in Sect. 5.11. What should be added is that our limited quantitative experiments do not sufficiently answer considerations regarding complex discourse quality metrics such as polarization into separate communities of users. So far, we have only simulated and analyzed one-turn communicative behavior. However, our framework also allows us to assess more sophisticated metrics that span discussion threads.

Also, to what extent computational models capture nuanced variations in argumentation styles across different user communities would require qualitative evaluations, which we have not included

in this paper. Discourse patterns can vary significantly between different communities, each with its own linguistic norms, interaction styles, and argumentation preferences. This heterogeneity presents both methodological and theoretical challenges. From a methodological perspective, we must consider whether our current modeling approaches are sufficiently sophisticated to capture these variations or if we need to develop more specialized, community-specific models. Theoretically, this opens up a broader question about the nature of social media discourse: Are social networks inherently so heterogeneous that meaningful modeling requires a community-by-community approach, rather than attempting to develop a universal model for online argumentation?

6.4 Potential Risks

As is typical for AI methods, the modeling approach presented in this paper is a dual-use technology. While social simulation - more concretely, replication of social networks and imitation of social media users - is primarly intended to be used to analyze social networks. However, the findings can also be used to adjust network mechanics, like ranking or filtering, which in turn influences public debate and opinion formation of users on those networks. Again, this can be used to improve debate quality and contribute to well-informed opinion formation. However, it can also be used to spread one-sided propaganda, misleading information, or manipulative advertisements. The reader should be aware of this negative potential.

References

Abubakar Abid, Maheen Farooqi, and James Zou. Persistent anti-muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pages 298–306, 2021. 27, 34

Lisa P Argyle, Ethan C Busby, Nancy Fulda, Joshua R Gubler, Christopher Rytting, and David Wingate. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3): 337–351, 2023. 25, 26, 27, 31, 34, 35

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. 42, 43

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference*

on Fairness, Accountability, and Transparency, FAccT '21, page 610–623, New York, NY, USA, 2021. Association for Computing Machinery. 27

Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, and Damon Woodard. Emulating author style: A feasibility study of prompt-enabled text stylization with off-the-shelf llms. In *Proceedings of the 1st Workshop on Personalization of Generative AI Systems (PERSONALIZE 2024)*, pages 76–82, 2024. 25

Axel Bruns. After the 'apicalypse': Social media platforms and their fight against critical scholarly research. *Disinformation and Data Lockdown on Social Platforms*, pages 14–36, 2021. 26

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023. 27

Bradley Carron-Arthur, John A Cunningham, and Kathleen M Griffiths. Describing the distribution of engagement in an internet support group by post frequency: A comparison of the 90-9-1 principle and zipf's law. *Internet Interventions*, 1(4):165–168, 2014. 38

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022. 28

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 41, 42

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36, 2024. 27

Jesse Graham, Jonathan Haidt, and Brian A Nosek. Liberals and conservatives rely on different sets of moral foundations. *Journal of personality and social psychology*, 96(5):1029, 2009. 26, 28, 29

Peter K Hatemi, Charles Crabtree, and Kevin B Smith. Ideology justifies morality: Political beliefs predict moral foundations. *American Journal of Political Science*, 63(4):788–806, 2019. 26

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021. 41

Basel Al-Sheikh Hussein. The sapir-whorf hypothesis today. *Theory and Practice in Language Studies*, 2(3):642–646, 2012. 28, 34

Ben Hutchinson, Vinodkumar Prabhakaran, Emily Denton, Kellie Webster, Yu Zhong, and Stephen Denuyl. Social biases in nlp models as barriers for persons with disabilities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5491–5501, 2020. 27, 34

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL `https://arxiv.org/abs/2310.06825`. 20, 21

Dominik K Kanbach, Louisa Heiduk, Georg Blueher, Maximilian Schreiter, and Alexander Lahmann. The genai is out of the bottle: generative artificial intelligence from a business model innovation perspective. *Review of Managerial Science*, 18(4):1189–1220, 2024. 27

Kisok R Kim, Je-Sang Kang, and Seongyi Yun. Moral intuitions and political orientation: Similarities and differences between south korea and the united states. *Psychological reports*, 111(1):173–185, 2012. 29

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022. 20

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*, 2023. 20

Chin-Yew Lin and Franz Josef Och. Orange: a method for evaluating automatic evaluation metrics for machine translation. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 501–507, 2004. 42

Ilya Loshchilov, Frank Hutter, et al. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 5, 2017. 41

Bolei Ma, Berk Yoztyurk, Anna-Carolina Haensch, Xinpeng Wang, Markus Herklotz, Frauke Kreuter, Barbara Plank, and Matthias Assenmacher. Algorithmic fidelity of large language models in generating synthetic german public opinions: A case study. *arXiv preprint arXiv:2412.13169*, 2024. 35

Robert W McGee. Is chat gpt biased against conservatives? an empirical study. *An Empirical Study (February 15, 2023)*, 2023. 30, 31

Paul D. O'Brien and Richard C. Nicol. Fipa—towards a standard for software agents. *BT Technology Journal*, 16:51–59, 1998. 34

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023. 15, 18, 27, 31, 34

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020. 20

David Rozado. The political biases of chatgpt. *Social Sciences*, 12(3):148, 2023. 23

Jérôme Rutinowski, Sven Franke, Jan Endendyk, Ina Dormuth, Moritz Roidl, and Markus Pauly. The self-perception and political biases of chatgpt. *Human Behavior and Emerging Technologies*, 2024(1): 7115633, 2024. 30, 31

Murray Shanahan. Simulacra as conscious exotica. *Inquiry*, pages 1–29, 2024. 27

Na Sun, Patrick Pei-Luen Rau, and Liang Ma. Understanding lurkers in online communities: A literature review. *Computers in Human Behavior*, 38:110–117, 2014. 38

Timm Teubner, Christoph M Flath, Christof Weinhardt, Wil van der Aalst, and Oliver Hinz. Welcome to the era of chatgpt et al. the prospects of large language models. *Business & Information Systems Engineering*, 65(2):95–101, 2023. 25

Alla Tiunova and Felipe Muñoz. Chatgpt: Using ai in social studies academic research. *Available at SSRN 4451612*, 2023. 25

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkwar, and Graham Neubig. Do llms exhibit human-like response biases? a case study in survey design. *Transactions of the Association for Computational Linguistics*, 12:1011–1026, 2024. 35

Petter Törnberg, Diliara Valeeva, Justus Uitermark, and Christopher Bail. Simulating social media using large language models to evaluate alternative news feed algorithms. *arXiv preprint arXiv:2310.05984*, 2023. 19

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 le language model based autonomous agents. *Frontiers of Computer Science*, 18(6):186345, 2024. 34

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021. 30

G Weiss. *Multiagent systems: a modern approach to distributed artificial intelligence*. MIT press, 1999. 33

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136, 2023. 19

Shicheng Xu, Liang Pang, Huawei Shen, and Xueqi Cheng. Nir-prompt: A multi-task generalized neural information retrieval training framework. *ACM Transactions on Information Systems*, 42(2):1–32, 2023. 25

Yu Yu, Chao-Han Huck Yang, Jari Kolehmainen, Prashanth G Shivakumar, Yile Gu, Sungho Ryu Roger Ren, Qi Luo, Aditya Gourav, I-Fan Chen, Yi-Chieh Liu, et al. Low-rank adaptation of le language model rescoring for parameter-efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023. 42

Xinyang Zhang, Yury Malkov, Omar Florez, Serim Park, Brian McWilliams, Jiawei Han, and Ahmed El-Kishky. Twhin-bert: A socially-enriched pre-trained language model for multilingual tweet representations at twitter. In *Proceedings of the 29th ACM SIGKDD conference on knowledge discovery and data mining*, pages 5597–5607, 2023. 41, 42, 43, 44

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