Social Intelligence Modeling: A Comprehensive Survey from Social Perception to Social Simulation

ZIKAI SONG*, Huazhong University of Science and Technology, China
XIAJIE LI*, Huazhong University of Science and Technology, China
YUNYAO ZHANG, Huazhong University of Science and Technology, China
XINGLANG ZHANG, Huazhong University of Science and Technology, China
WEI YANG, Huazhong University of Science and Technology, China
JUNQING YU, Huazhong University of Science and Technology, China

Social systems shape communication, economic activity, governance, and everyday life, while artificial intelligence increasingly changes how these systems are observed, interpreted, modeled, and influenced. Yet relevant research remains fragmented across artificial intelligence, human-computer interaction, communication, and social sciences, spanning heterogeneous tasks and rapidly evolving paradigms. To organize this landscape, we introduce **Social Intelligence Modeling (SIM)** as a comprehensive perspective on how AI perceives social signals, models individual and collective behavior, generates socially situated outputs, and simulates social-system evolution. We curated roughly 1,000 papers spanning over 20 tasks, with around 300 analyzed in detail[1]. We organize the field into four progressively integrated levels: **Perception**, which identifies social signals and states; **Understanding**, which models structures, relations, and behavioral patterns; **Generation**, which produces socially meaningful content and interventions; and **Simulation**, which reconstructs dynamic social processes and explores possible futures. Particular emphasis is placed on social simulation as an emerging core frontier of SIM, for which we systematically examine its paradigms, scales, application and broader societal considerations. By synthesizing fragmented research, this survey lays the groundwork for understanding and shaping AI-mediated societies.

CCS Concepts: • **Computing methodologies → Artificial intelligence**; **Modeling and simulation**; • **Applied computing → Law, social and behavioral sciences**; • **Human-centered computing → Collaborative and social computing**.

Additional Key Words and Phrases: social intelligence modeling, social simulation, large language models, computational social science, multi-agent systems, collective behavior

1 Introduction

1.1 Background

Social systems are fundamental to human life, shaping how people communicate, participate in economic activity, access public services, and organize everyday social relations. ***Artificial intelligence*** has become increasingly intertwined with these systems, supporting the analysis of human emotions and opinions, the creation and moderation of social content, and data-informed decision-making in areas such as public governance, commerce, and social services. We define **Social Intelligence Modeling (SIM)** as the study of representing, understanding, generating, and simulating human and collective behaviors in digital social systems using AI techniques, especially large language models and

*Both authors contributed equally to this work.
[1]The complete paper list is in Project Page: https://github.com/sait-crypto/Awesome-Social-Intelligence-Modeling-System

Authors' Contact Information: Zikai Song, Huazhong University of Science and Technology, School of Computer Science and Technology, Wuhan, China, skyesong@hust.edu.cn; Xiajie Li, Huazhong University of Science and Technology, School of Cyber Science and Engineering, Wuhan, China, xiajie2026@hust.edu.cn; Yunyao Zhang, Huazhong University of Science and Technology, School of Cyber Science and Engineering, Wuhan, China,; Xinglang Zhang, Huazhong University of Science and Technology, School of Cyber Science and Engineering, Wuhan, China,; Wei Yang, Huazhong University of Science and Technology, School of Cyber Science and Engineering, Wuhan, China,; Junqing Yu, Huazhong University of Science and Technology, School of Cyber Science and Engineering, Wuhan, China,

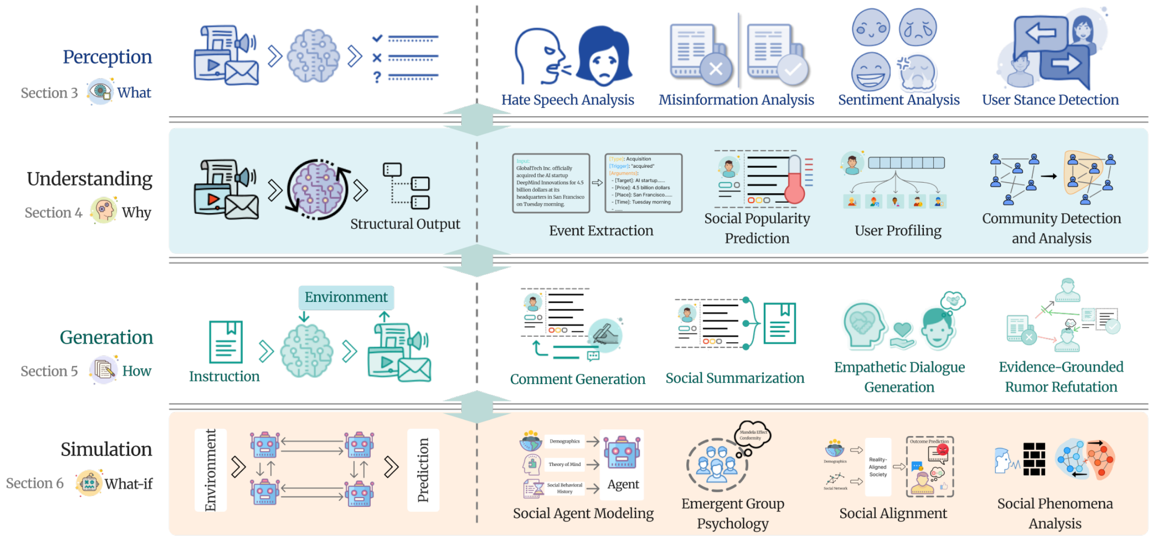

Fig. 1. Four progressive cognitive levels of Social Intelligence Modeling (SIM), evolving from perception to simulation. SIM forms a cognitive hierarchy that advances from recognizing social signals (**Perception**), to understanding social structures and behaviors(**Understanding**), to generating meaningful interactions and interventions (**Generation**), and ultimately to simulating dynamic social processes and anticipating possible futures (**Simulation**).

multi-agent frameworks. This survey systematically reviews AI techniques for social systems, and introduces **SIM** as a unified framework for organizing how AI perceives social signals, models individual and collective behavior, generates socially situated actions, and simulates the evolution of social systems.

1.2 Challenges

A comprehensive understanding of SIM is inherently challenging for two reasons: (1) the ***subject area*** spans multiple disciplines, including artificial intelligence, human–computer interaction, sociology, communication, and computational social science; (2) it also draws on a broad range of ***technical paradigms***, from multimodal representation and semantic reasoning to generative modeling and agent-based systems.

Challenge 1: Fragmented and heterogeneous task landscape. SIM encompasses a wide spectrum of tasks with substantially different objectives, outputs, and modeling assumptions. ***Established tasks*** such as sentiment analysis, misinformation detection, hate-speech detection play fundamental roles in interpreting social content; ***less conventional tasks***, including empathetic response generation, evidence-grounded rumor refutation, event discovery, social summarization, and intervention-oriented generation, are also becoming increasingly important; Meanwhile, ***emerging directions*** such as social agent modeling, collective behavior prediction, and large-scale social simulation continue to extend the boundaries of SIM. The coexistence of these diverse and rapidly evolving tasks makes it difficult to identify their intrinsic relationships, or organize them within a coherent and sufficiently inclusive taxonomy.

Challenge 2: Dispersed and difficult-to-access research knowledge. The breadth and fragmentation of SIM make the field difficult to navigate for both newcomers and experienced researchers. First, ***inconsistent terminology*** prevents effective literature retrieval: searches based on individual task names or technical keywords often capture only narrow segments of the field, making it difficult for newcomers to obtain a comprehensive overview or recognize the connections among related research directions. Second, ***related studies are widely distributed*** across AI, HCI, computational social science, communication, and various domain-specific venues, further increasing the difficulty of systematic discovery and tracking. Social simulation is a representative example, as closely related studies are often published under different terminologies, including agent-based modeling, LLM-based agents, multi-agent systems, and society simulation, making both foundational work and recent advances difficult to identify, compare, and follow.

1.3 Analysis of Existing Surveys

We compare the existing surveys with our work in Table 1. Existing surveys either focus on individual subdomains [150, 152, 188, 254], summarize earlier machine-learning or generative perspectives on social media analysis [136, 213], or discuss broader intersections among AI, computational social science, and LLM-based agents [141, 211, 245].

However, existing surveys provide insights into specific tasks or communities, yet remain fragmented without a unified SIM perspective. Many were developed before the rise of LLMs, multimodal foundation models, and intelligent agents, thus overlooking recent paradigm shifts. More importantly, the absence of a clear definition and taxonomy of SIM prevents existing works from systematically organizing its modeling objects, tasks, and methodologies. As a result, current surveys remain limited within individual domains and lack a comprehensive understanding of SIM as an emerging interdisciplinary field.

1.4 Contributions

We present a comprehensive and structured survey of SIM that clarifies its conceptual boundaries, organizes its internal dependencies, and provides a systematic literature map and forward-looking research agenda.

1) Comprehensive coverage. We conduct a broad review of more than ***1,000 papers*** spanning over ***20+ tasks***, covering the major research problems, methodological paradigms, datasets, evaluation settings, and application scenarios [2]. This coverage brings together research that has previously been scattered across separate communities and task-specific surveys. To the best of our knowledge, this is the most extensive survey of SIM to date in terms of both the number of studies examined and the breadth of tasks covered.

2) A coherent four-level taxonomy. Inspired by human cognitive processes, we organize SIM into a progressive path (Figure 1): **Perception, Understanding, Generation,** and **Simulation**. It traces a natural progression from identifying ***what*** is occurring, to explaining ***why*** it occurs, determining ***how*** to respond, and exploring ***what would happen if*** particular social conditions or interventions were introduced.

3) A dedicated focus on the core frontier of SIM—Social Simulation. We place particular emphasis on ***social simulation***, which is becoming a central frontier of SIM. Traditional tasks primarily describe or interpret observed social data, whereas simulation enables researchers to reconstruct social processes, forecast evolution, and test interventions before they are deployed in the real world. Such capabilities are increasingly important but remain insufficiently covered by existing surveys. We provide a dedicated synthesis of this direction, including methodologies, simulation scales, applications, and broader considerations.

4) An in-depth analysis of challenges, future directions, and societal implications. We further identify ***six fundamental challenges*** that will shape the future development of SIM. These challenges expose limitations that cannot be resolved solely by scaling existing models. Building on these challenges, we outline ***future connections***

[2] We select approximately 300 influential works for detailed discussion and citation in the survey, full collected papers released in the Project page: https://github.com/sait-crypto/Awesome-Social-Intelligence-Modeling-System.

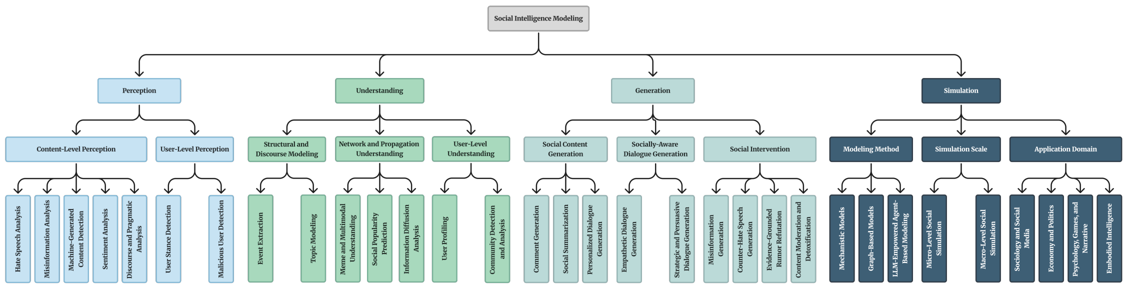

Fig. 2. Hierarchical taxonomy of SIM. Perception, Understanding, and Generation are organized into task families, while Simulation is characterized by modeling method, scale, and application domain.

Table 1. Comparison between our survey and representative related surveys.

Work	Year	Focus	Social Scope		SIM Modeling Objects				Paradigm and Support			
			SM	CSS	Content	Actors	Relations	Evolution	LLM/FM	Agent	Multi.	Res./Eval.
T.K. et al. [213]	2021	SM	●	◐	●	◐	◐	◐	×	×	◐	◐
Matwin et al. [136]	2021	SM	●	◐	●	◐	●	●	×	×	◐	◐
Xu et al. [245]	2024	CSS	◐	●	◐	●	◐	◐	✓	✓	○	●
Thapa et al. [211]	2025	CSS	◐	●	●	◐	◐	◐	✓	×	◐	◐
Mou et al. [141]	2026	SS	○	●	◐	●	●	●	✓	✓	○	●
Our Survey	2026	SIM	●	●	●	●	●	●	✓	✓	●	●

●: broadly covered; ◐: partially covered; ○: not a major focus.
SM = social media; CSS = computational social science; SS = social simulation; SIM = social intelligence modeling; FM = foundation model.
Content, Actors, Relations, and Evolution correspond to the major modeling objects in the SIM environment. Res./Eval. covers datasets, benchmarks, or systematic evaluation discussions. The table characterizes related surveys by their stated focus and SIM-relevant modeling objects.

between SIM and social theory, world models, embodied intelligence, and human–AI coexistence, and examine how SIM may influence the design, governance, and understanding of increasingly AI-mediated societies.

2 Scope and Overview

2.1 Modeling Objects

We organize online social environments around four modeling objects:

- ***Social content***, which captures the information, emotions, stances, intentions, and cultural signals expressed through posts, comments, videos, and other social artifacts;
- ***Social actors***, including users, organizations, communities, and bots that perceive environments, make decisions, interact, and exert influence;
- ***Interaction networks***, which represent relational structures such as following, replying, reposting, and co-exposure that govern information diffusion and social influence;
- ***Temporal evolution***, which models the dynamic changes of content, actors, networks, opinions, and collective behaviors over time.

These objects appear across representative SIM tasks: misinformation analysis links claims, sources, propagation, and correction; sensitive discourse requires cultural and conversational grounding; actor-centered tasks model users through histories, preferences, and networks; and social simulation integrates agents, interactive environments, and temporal dynamics for counterfactual analysis.

2.2 Positioning Social Intelligence Modeling

The emergence of SIM follows the evolution of AI from task-specific intelligence toward general-purpose social intelligence. Early studies relied on handcrafted features and statistical models to analyze specific social signals, such as sentiment, misinformation, and user behaviors. Representation learning [192, 194] and Transformer networks [129, 193, 195] further enabled semantic and structural modeling through pre-trained language models and graph neural networks [196]. With the rise of LLMs [111, 112, 285], multimodal foundation models [194], and intelligent agents [291], AI systems have gradually evolved from recognizing social information to reasoning, generating, and simulating social behaviors. SIM can therefore be viewed as an integrated framework that connects these technological advances for modeling complex social systems.

Perception focuses on identifying and extracting fundamental social signals from multimodal social data. It is closely related to existing tasks in sentiment analysis, emotion recognition, stance detection, hate speech detection, misinformation detection, and multimodal content understanding. These tasks provide the basic capability to perceive opinions, emotions, intentions, and credibility signals expressed in social environments.

Understanding aims to interpret the underlying structures, behaviors, and mechanisms of social systems. It is closely connected to computational social science and social network analysis, including tasks such as event understanding, user profiling, community analysis, influence modeling, and popularity prediction. These studies investigate how individuals behave, how groups form, and how information propagates in social environments.

Generation focuses on creating socially meaningful content, responses, and interventions, connecting social computing with recent advances in generative AI. It relates to tasks such as dialogue generation, comment generation, social summarization, empathetic response generation, and rumor clarification. While general generative models focus on producing fluent and coherent outputs, SIM generation emphasizes the social context behind generation, including user preferences, interaction history, community norms, and potential social impact.

Simulation represents the integration of perception, understanding, and generation for modeling dynamic social processes. It is closely related to traditional computational social simulation and agent-based modeling, where artificial agents interact according to predefined rules to study collective behaviors. Recent LLM-based agents further enhance this paradigm by introducing memory, reasoning, planning, and natural communication abilities.

2.3 Organization and Taxonomy

Based on the above discussions„ our taxonomy is organized around four levels: Perception, Understanding, Generation, and Simulation. Sections 3–6 review representative methods and tasks within each level, while Section 7 discusses open challenges and future directions of SIM.

3 Perception: Identifying Social Signals

Perception represents the entry level of social intelligence modeling. It identifies socially meaningful signals in content and actors, including harmfulness, credibility, sentiment, stance, provenance, and maliciousness. In SIM, these outputs are not self-contained labels but situated judgments whose meaning depends on context, norms, modality, actor history, and temporal dynamics. They provide the evidence base for subsequent Understanding, Generation, and Simulation.

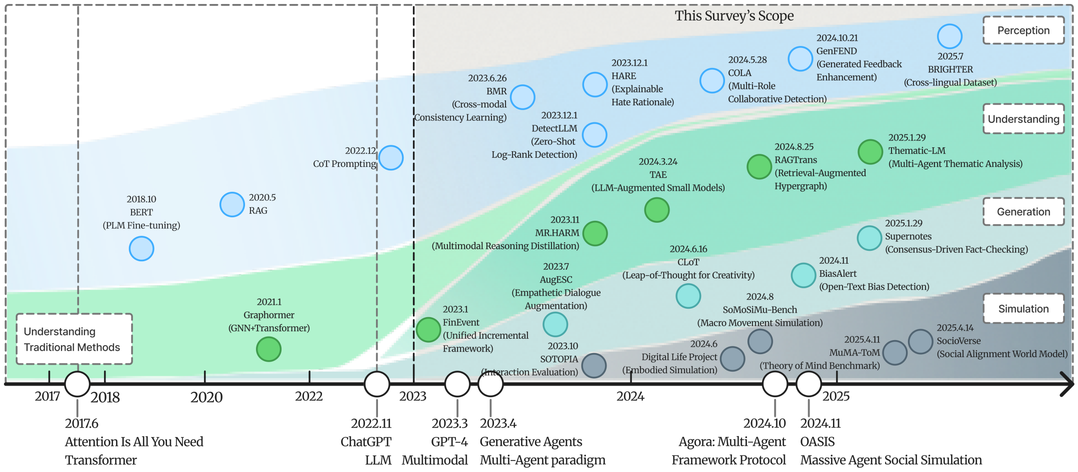

Fig. 3. Roadmap of AI-driven social intelligence modeling. The shaded region indicates the main scope of this survey. Circular markers denote influential studies within their corresponding paradigms. Under the LLM paradigm, understanding tasks shift fundamentally from statistical prediction to structured reasoning and social generation, so we split the two stages.

3.1 Content-Level Perception

Content-level perception treats posts, comments, replies, titles, descriptions, images, memes, and short videos as the primary units of analysis, while incorporating actor histories, networks, and temporal context when required for interpretation.

3.1.1 Hate Speech Analysis. Hate speech analysis shows why social perception cannot be reduced to surface classification. Although its immediate output is often a harmfulness label, the judgment depends on implication, target identity, modality interaction, and community norms. Its cultural dependence challenges the assumption that such labels constitute stable ground truth [66, 214]. ***Implicit and pragmatic signals.*** Harmful intent is increasingly conveyed through dogwhistles, metaphors, coded references, and context-dependent expressions [94, 138]. Such signals can mislead even strong LLMs [270], motivating methods that incorporate conversational context and dialogue structure [16, 59]. ***Multimodal and perspectival judgment.*** In memes and visual posts, harmful meaning often emerges from image–text interaction rather than either modality alone. Recent methods use optimal-transport alignment to connect cross-modal cues [278, 279], hard-negative learning to distinguish misleading associations [137], and LLM-guided reasoning to recover hidden targets and emotions [97]. Beyond recognition, explanation-oriented work generates hate rationales [255] and assists moderation [23]. Multi-agent debate may further help study how heterogeneous communities interpret harmful speech and respond to moderation [162].

3.1.2 Misinformation Analysis. Misinformation is not a static content property but a process through which claims are created, circulated, contested, reframed, and corrected. We therefore organize the literature around a lifecycle of creation, propagation, and post-propagation. ***(1) Creation:*** Research examines rhetorical manipulation and propaganda [68, 85], inconsistent event relations [103], and multimodal fabrication [217, 263]. Reliable perception therefore depends on structured relations among claims, events, sources, and modalities rather than detector architecture alone. ***(2)***

Propagation: During propagation, the task shifts from isolated content judgment to evidence-aware and network-aware verification. Retrieval- and knowledge-augmented methods connect claims to external evidence and entity relations [42, 130, 153]. Evidence that rumor trees are often "shallow and wide" challenges assumptions behind deep propagation models [40], motivating zero-shot detection for sparse emerging events [119]. These approaches remain constrained by incomplete evidence, topology, and early observations. ***(3) Post-propagation and correction:*** After broad diffusion, perception must account for susceptibility, emotional co-evolution, exposure effects, and correction [124, 183]. Because many reactions and alternative paths remain unobserved, recent work uses generative simulation to augment perception, including silent-user simulation [147], unobserved-path forecasting [281], and adversarial natural-language learning [212].

3.1.3 Machine-Generated Content Detection. Machine-generated content (MGC) detection determines whether a social signal is human-produced, machine-produced, or hybrid. Its role in SIM is provenance protection: synthetic posts can distort evidence, intent, and apparent public opinion while contaminating actor profiles, retrieval evidence, diffusion traces, and simulation inputs. Social media further weakens conventional authorship cues because content is short, personalized, multilingual, multimodal, and culturally situated [132, 180]. Training-free methods exploit generation artifacts to provide efficient provenance signals [199, 246], while supervised semantic detectors may improve cross-domain generalization but require continual data renewal [116]. The resulting trade-off is efficiency versus adaptability. Social settings also require finer judgments about AI involvement, co-writing, and downstream circulation [35]. Once synthetic content circulates, static detectors face adaptive evasion [44] and polluted retrieval evidence [216]. MGC detection thus exposes a broader SIM problem: the observed social environment may itself be partially synthetic. Robust provenance modeling should account for interacting generators, detectors, users, and propagation environments rather than treating generated content as isolated text; multi-agent debate offers one early approach [127].

3.1.4 Sentiment Analysis. Sentiment analysis assigns affective labels or distributions, but its target is not an objective emotional fact. The "Sentiment Problem" highlights the fragility of sentiment labels across domains, communities, and situations [215]. SIM therefore treats sentiment as situated affect: how emotions are expressed, interpreted, and contested within cultural, conversational, multimodal, and temporal contexts. ***Perspectival affect.*** Ambiguous annotations may represent distributions of viewpoints rather than annotation noise [235], while LLMs may reproduce gender and nationality biases in emotion attribution [9, 167]. These findings call for perspectival and community-specific alignment rather than merely larger encoders. Affective labels should therefore be interpreted as signals about how particular actors or communities perceive an event, claim, or interaction. ***Granular and interactional affect.*** Reliable analysis must bind affect to the correct target, modality, and conversational state. Aspect-based methods model structured relations among targets, aspects, opinions, and emotions [128], while multimodal approaches align textual, visual, and incomplete evidence [32, 232]. Coupled Mamba [113] captures cross-modal dynamics via coupled state modeling. Difficult cases include metaphorical image–text relations [280] and conversational emotion that changes with speaker state and interaction history [80, 225]. Reader-agent frameworks can estimate latent reactions [118], but such responses remain modeling assumptions requiring validation against observed users. Sentiment analysis therefore moves beyond polarity classification toward situated affect interpretation. Even strong emotion classifiers may fail when literal expression diverges from intended or perceived meaning, as in sarcastic praise or strategic skepticism. This limitation leads directly to discourse and pragmatic analysis.

3.1.5 Discourse and Pragmatic Analysis. Discourse and pragmatic analysis lies between Perception and Understanding: while it often produces labels such as sarcasm, humor, or manipulation, their recognition depends on deeper modeling of speaker intent, cultural conventions, conversational context, and social relations. ***Micro-level pragmatic expressions.*** Sarcasm commonly depends on incongruity among literal expression, factual context, emotion, and speaker intention [139, 174]. Multimodal methods further model incongruity across image and text [31, 173]. Related phenomena are less stable across cultures and communities. Humblebragging disguises self-promotion as complaint or modesty [146], while euphemistic meanings evolve rapidly and require locally grounded interpretation [74, 114]. These labels are relations among expression, speaker, audience, and shared background rather than intrinsic properties of text. ***Macro-level discourse analysis.*** Macro-level discourse analysis examines how pragmatic expression functions as a tool for ideology, persuasion, or narrative disruption. Dogwhistles encode group-specific meanings invisible to outsiders [138] and resemble culturally grounded word-sense disambiguation [94]. Whataboutism redirects blame and derails argumentation [165], while mental manipulation exploits interpersonal vulnerability [226]. These tasks require identifying what an utterance does within an interaction, not merely what it states. ***Pragmatic judgment is also shaped by social structure***. Situational context can invert perceived offensiveness [303], social relationships constrain appropriateness [84], and conversation kernels expose dependencies across speakers and turns [1]. The relevant context may also be hidden in subcultural knowledge or conversation history, making pragmatic labels especially unstable across communities.

3.2 User-Level Perception

User-level perception shifts the unit of analysis from content to actors. Profiles, posting histories, interactions, and network positions provide partial estimates of actor states and bridge content perception with behavioral modeling.

3.2.1 User Stance Detection. User stance detection infers an actor's attitude toward a target. Although its immediate output is typically a favor, against, or neutral label, stance is often implicit in sarcasm, domain knowledge, and interaction history, making surface text insufficient. ***Knowledge augmentation*** grounds implicit opinions [105], explicit reasoning exposes intermediate inference [206], and multi-perspective verification mitigates unsupported stance inferences under weak evidence [292]. These methods improve semantic coverage but often assume that targets and contexts are known in advance. ***Real social settings*** further involve unknown or evolving targets, long conversations, domain shifts, and temporal drift. Research has therefore moved toward open-target extraction [115], conversational opinion tracking [135], synthetic-data-assisted adaptation [293], and modeling stance under temporal drift [143]. Stance reflects values, identities, roles, and temporal interactions rather than a stable textual property. Collaborative agents [99] and value-distribution models [86, 277] extend this direction toward latent beliefs and unseen topics.

3.2.2 Malicious User Detection. Malicious user detection identifies actors whose coordinated or adaptive behavior harms social environments. Its evidence is distributed across language, temporal activity, behavioral consistency, and network position, making isolated-post classification insufficient. Bots and coordinated accounts can influence public events and discourse [204]. Graph methods capture heterogeneous social relations [29], but strategic actors may evade detection through heterophilic links with humans [171]. ***Recent methods*** therefore move from node classification toward semantic–structural actor modeling, combining semantic and pragmatic signals [50] with graph topology [110]. This shift reflects that maliciousness is neither solely a content property nor a graph anomaly, but a behavioral state defined by coordination, consistency, temporal trajectory, and content–network interaction. ***Simulation*** further exposes distribution shift because malicious users are adaptive social actors. BotSim uses LLM-based agent-based modeling to generate malicious botnets embedded in social environments, under which conventional detectors degrade [170]. This degradation suggests that detectors fitted to observed behavior may not generalize to strategically generated actors. Malicious user detection therefore links perception with simulation, where adversarial and defensive behaviors must be modeled as interacting and adaptive processes.

4 Understanding: Modeling Social Structures

Understanding transforms perceived signals into structured social states spanning events, themes, actors, communities, networks, multimodal meaning, and temporal dynamics. Rather than producing isolated labels, it organizes social evidence into representations that explain how content, actors, and relations form coherent social situations. These states provide the explanatory context, behavioral priors, and validation targets required by Generation and Simulation.

4.1 Structural and Discourse Modeling

Structural and discourse modeling turns noisy social streams into event structures, thematic organization, and cross-modal social meaning. Its goal is not merely to aggregate similar content, but to reconstruct interpretable social states that support explanation, intervention, and simulation.

4.1.1 Event Extraction. Event extraction reconstructs events, participants, arguments, and relations from fragmented social media streams. Its output is an event-centered social state describing what happened, who was involved, and how events are connected. Such states provide a temporal and causal substrate for later reasoning rather than only a collection of extracted facts. Event discovery identifies evolving boundaries in dynamic or cross-lingual settings [163], while hypergraph-based methods capture changing event structure [267]. LLM-based extraction further supports multilingual event monitoring [20]. MA-VLAD [49] models fine-grained event dynamics with attention aggregation. Argument extraction grounds events in actors, locations, objects, and evidence, including long-tail arguments [28] and multimedia arguments [69, 179]. Relation extraction then connects events through causal or interactional dependencies using zero-shot reasoning [78] or teacher–student alignment [121].

4.1.2 Topic Modeling. Topic modeling transforms fragmented discourse into interpretable themes, collective attention states, and community frames. It belongs to Understanding because its output is not a content label but an organized account of what a population is discussing, how themes relate, and how collective attention is distributed. ***Relational models*** recover conversational context [231], while platform-invariant methods align themes across environments [93] and multilingual methods extend them across languages [4]. These directions move topic analysis by incorporating interaction, platform variation, and linguistic diversity. ***Recent work*** further shifts from flat clusters toward structured thematic knowledge. LLMs induce interpretable taxonomies [274], hierarchical methods recover level-wise news themes [67], and multi-agent experts support qualitative interpretation [172]. The key question is whether the identified themes capture real social meanings rather than patterns learned by the model.

4.1.3 Meme and Multimodal Understanding. Meme and multimodal understanding reconstructs social meaning that emerges across text, images, templates, and community knowledge. Unlike perception tasks that assign labels to multimodal inputs, it organizes signals into entities, visual roles, metaphors, contradictions, and cultural frames. Cross-modal grounding links textual entities with visual regions [265], while span-based methods reduce alignment noise [247]. Because meaning may also depend on metaphor or missing context, other methods recover latent image–text relations [280] or retrieve background knowledge [101]. Explanation-oriented systems further clarify meme background [182] and hidden visual implications [181]. SF2̂T [75] improves Video-LLMs' fine-grained understanding via self-supervised fragment tuning. MA-PDM [299] enables fine-grained anomaly detection with appearance-motion guided diffusion. CurEvo [269] enhances video understanding through curriculum-guided self-evolution. Current LMMs remain weaker than humans on meme semantics [83] and socially situated context [257]. Moreover, static image–text pairs omit the imitation, variation, and community-level evolution central to memetics [152].

4.2 Network and Propagation Understanding

Network and propagation understanding explains how content becomes visible and spreads through topology, temporal traces, exposure, and engagement. It infers relational and temporal regularities from observed processes, whereas bottom-up execution of counterfactual dynamics belongs to Simulation.

4.2.1 Social Popularity Prediction. Social popularity prediction estimates the engagement potential or future trajectory of content from its attributes, actor profiles, platform exposure, early feedback, and time. Popularity is not an intrinsic property of content, but a platform-mediated attention state emerging from the interaction among content, users, exposure mechanisms, and prior reactions. Existing work captures complementary sources of evidence. Pre-diffusion models [242] combine content and user signals. Retrieval-augmented methods [34] compare posts with historically similar diffusion cases. Lifecycle forecasting [248] estimates later trajectories from early engagement. STAP [220] introduces spatio-temporal enlargement for long-range video perception and scalable popularity prediction. OmniTrend [261] decouples content attractiveness and exposure effects for interpretable popularity modeling. MVP [262] leverages multimodal features and contextual signals for video popularity prediction, winning the SMP 2025 Video Track. These methods learn popularity patterns from the past, but struggle to explain how popularity evolves under changing social environments.

4.2.2 Information Diffusion Analysis. Information diffusion analysis recovers propagation regularities from observed cascades. It seeks to explain how network structure, exposure, temporal ordering, and actor characteristics shape the movement of information. ***Classical models*** represent spread as stochastic transitions over predefined networks [188], offering interpretability but often simplifying semantic context, hidden exposure, and cognitive heterogeneity. ***Modern approaches*** enrich visible topology with latent and higher-order structure. Latent influence networks infer unobserved exposure relations [63]. Temporal hypergraphs and multiscale representations capture cascade dynamics across macro diffusion patterns and micro user relations [81, 241]. Identity-aware modeling further shows that diffusion is not only transmission over edges, but also a socio-cognitive process shaped by actor identity and group affiliation [2]. These approaches provide empirical propagation priors, but they remain dependent on historical cascades and may fail when interventions alter exposure. ***Diffusion understanding*** estimates propagation patterns and influence structures from observed cascades, supporting intervention design through influence optimization [102, 240]. ***Propagation synthesis*** begins to explore possible futures [281], but Simulation further models adaptive agents and changing environments beyond fitted diffusion assumptions.

4.3 User-Level Understanding

User-level understanding models individual heterogeneity from digital footprints. Rather than treating users as interchangeable network nodes, it constructs actor states from preferences, identities, values, motives, and behavioral histories. These states explain why actors exposed to similar content may interpret, share, or respond to it differently.

4.3.1 User Profiling. User profiling aggregates fragmented posts, conversations, and other traces into structured actor representations. Open-domain profiling moves beyond explicit demographic labels toward richer states inferred from unstructured conversation [230]. Such profiles may encode preferences, interests, values, communication patterns, and other behavioral tendencies needed by personalized generation and heterogeneous simulation. ***The main challenge is to*** summarize long, sparse, and temporally ordered histories into states that remain stable but adaptive. Dynamic graphs capture latent relations among posts [253]. Text-graph pretraining supports long histories and missing network information [233], while continual user modeling preserves representations as tasks evolve [90]. Together, these methods move user modeling from independent post aggregation toward structured and temporally evolving actor states. These profiles provide behavioral priors for community analysis, personalized dialogue, and heterogeneous simulation.

4.3.2 Community Detection and Analysis. Community detection converts actor relations into meso-level social structure. Unlike topic modeling, which organizes content, it groups actors through shared ties, attributes, interests, ideologies, or practices [17]. ***Structural methods*** improve discovery under sparse, noisy, incomplete, and large-scale graphs [238], while attribute-aware methods address missing node information [260]. Other work infers latent communities from ideology and behavior. Stance-aware embeddings reveal ideological groups and differences in susceptibility to low-credibility information [10], while LLM-based analysis identifies shared community practices [87]. Human–AI networks further require models that account for embedded AI actors while preserving human connectivity [36]. Community structures connect individual actors with collective behavior. They support analysis of influence, norms, vulnerability, and human–AI interaction, while providing meso-level environments in which diffusion, coordination, and intervention can later be simulated.

5 Generation: Producing Socially Situated Actions

Generation is the first SIM level that acts on, rather than only analyzes, the social environment. Perception identifies signals, Understanding organizes them into social states, and Generation asks how to respond. It concerns socially situated action rather than open-ended text production: generated outputs can reshape user reactions, platform exposure, trust, attention, beliefs, and downstream social dynamics.

5.1 Social Content Generation

Social content generation produces platform-native units such as comments and summaries. Because these outputs circulate through engagement mechanisms, the central question is not only whether they are fluent or relevant, but whether they function responsibly within platform feedback loops.

5.1.1 Comment Generation. Comment generation is a micro-level social action that may amplify attention, invite agreement, provoke conflict, or shift the tone of a discussion. Its main difficulty is that engagement signals such as likes are noisy preference proxies shaped by posting time, author influence, exposure, and platform bias. ***Recent work*** uses social feedback as a reward signal. PopALM aligns generation with popularity signals [264], while ViCo separates engagement from temporal exposure effects [201]. HotComment [236] models comment popularity through multimodal evaluation, prediction, and user behavior simulation. These methods improve social responsiveness, but optimizing short-term engagement may produce repetitive, sensational, or polarizing comments.

5.1.2 Social Summarization. Social summarization turns posts, memes, conversations, and long discussion threads into readable interpretive artifacts. Unlike ordinary condensation, it selects and organizes competing social signals and therefore depends on discourse structure, viewpoints, modalities, and temporal evolution. ***Recent work*** models conversational topology [57], sociological and semantic relations [295], multimodal context [96], and temporal evolution [189]. Structured outputs further support aspect-decomposed opinion comparison [109] and table-of-contents-style navigation through long or polarized discussions [203]. Summarization thus transforms structured social states into communicative artifacts, but also introduces framing risk by privileging some relations, emotions, or groups over others. ***Generated social content*** may also reproduce political bias [8], narrow thematic diversity [229], and flatten plural values [197]. These are cross-cutting constraints on socially situated generation because generated outputs may later become user-facing information, retrieval evidence, or behavioral priors for simulation.

5.2 Socially-Aware Dialogue Generation

Socially-aware dialogue moves Generation from one-way content production to interactive social action. It uses actor, context, and emotion models from Understanding to maintain identity, build trust, adapt to users, and pursue social goals.

5.2.1 Personalized Dialogue Generation. Personalized dialogue gives agents stable but adaptive social identities. Existing work constructs personas from dense descriptions and sparse attributes [205], customizes social behavior and inner traits across long interactions [300], and models speaker–addressee personas in multi-party dialogue [133]. The central issue is not persona richness alone, but whether identity remains credible as roles, relationships, and conversational goals change. Personalized dialogue therefore turns user modeling into behavior for interactive agents.

5.2.2 Empathetic Dialogue Generation. Empathetic dialogue adapts responses to users' emotional states. This matters in moderation, counseling, rumor correction, and counterspeech, where factually appropriate responses may fail when they ignore emotion. Current methods address data scarcity through LLM-based augmentation [296], regulate emotional flow through commonsense reasoning and reinforcement learning [21], and extend empathy to multimodal embodied interaction [48]. Empathetic dialogue is therefore relational action: it shapes trust, emotional alignment, and the conditions under which later interventions may succeed.

5.2.3 Strategic and Persuasive Dialogue Generation. Strategic dialogue extends identity and empathy toward goal-directed social interaction. Systems learn persuasive strategies through intent-to-strategy reasoning and simulation-based preference optimization [82], or improve factual grounding through retrieval-augmented zero-shot persuasion [53]. The central trade-off is effectiveness versus control: stronger persuasion may increase manipulation risk when factuality, intent, or user vulnerability is weakly governed. Generated conversations may also contain hidden harms and cultural biases, especially outside Western contexts [41], while multicultural norm extraction offers one route toward situated alignment [52]. Because norms remain subjective and contested, persuasive success alone does not establish socially beneficial interaction. Simulation can help examine how dialogue actions reshape trust, beliefs, behavior, and discourse before real-world intervention.

5.3 Social Intervention

Social intervention is the most action-oriented form of Generation. Models produce adversarial or corrective actions intended to change discourse dynamics, shifting evaluation from isolated output quality toward policy effects and downstream social consequences.

5.3.1 Misinformation Generation. Misinformation generation models AI-generated harm as an adaptive process rather than a static falsehood. Process-oriented work simulates stepwise narrative distortion through multi-agent interaction [125], while adversarial frameworks allow generators and detectors to co-evolve in natural language [212]. These systems treat deception as an evolving intervention that may reshape attention, beliefs, and defensive behavior while stress-testing detector robustness.

5.3.2 Counter-Hate Speech Generation. Counter-hate speech generation represents the corrective side of intervention. Recent work constrains generation by intended outcomes [70] and grounds counter-narratives in evidence [221]. Effectiveness and safety may nevertheless conflict: strong guardrails can make responses too weak to argue persuasively [13]. These systems therefore define candidate interventions whose persuasive impact, factuality, tone, and harm constraints must be considered together.

5.3.3 Evidence-Grounded Rumor Refutation. Evidence-grounded rumor refutation turns verification into corrective social action. In decentralized platforms, models must not only identify false claims, but also generate explanations that users can trust. Consensus-oriented systems synthesize diverse community notes into higher-agreement refutations [45], while joint stance–verification modeling connects factual judgment with audience position [252]. Delivery also matters: emotional framing can affect perceived credibility and diffusion [178], while value-tailored explanations may improve perceived helpfulness in simulated evaluation [54]. Rumor refutation therefore moves from one-to-many debunking toward trust-sensitive intervention grounded in evidence, audience state, and platform context.

5.3.4 Content Moderation and Detoxification. Content moderation is shifting from binary filtering toward generation-assisted governance. Detoxification rewrites harmful text while preserving meaning [168]; policy-aware moderation adapts to evolving policy rules [145]; explanation-based moderation exposes implied social biases for deliberative review [287]; and harm-aware re-ranking reduces harmful exposure without direct deletion [156]. The scope also includes privacy abstraction through rewriting sensitive self-disclosure [47]. These interventions shape what users can say, what others see, and how harmful exposure is transformed. Their central trade-off is control versus distortion, because moderation may reduce harm while suppressing context, imposing platform norms, or hiding disagreement.

6 Simulation: Constructing Social Evolution

Simulation is the most integrative level of SIM. It asks what would happen if social conditions, actors, or interventions changed, and closes the loop among Perception, Understanding, and Generation by modeling how signals, actors, interactions, generated actions, and temporal dynamics co-evolve in a computational environment. Table 3 provides a structured comparison of representative social simulation studies.

6.1 Methodological Paradigms: Mechanisms, Graphs, and Agents

The recent research of social simulation primarily follows three complementary paradigms: mechanistic models, graph-based models, and LLM-empowered agent-based modeling, as illustrated in Figure 4. They represent social evolution primarily through transition rules, network structures, or autonomous actors.

6.1.1 Mechanistic models. Mechanistic models formalize social dynamics through interpretable rules and continuous processes. Epidemiological models such as IC and SIR support analysis of threshold behavior and influence bounds [51, 122]. Opinion-dynamics models such as Friedkin–Johnsen enable polarization reduction and network intervention [95, 200], while temporal point processes, especially Hawkes processes, capture continuous-time events

Table 2. Representative Benchmarks and Datasets Across the Four Cognitive Levels

Task	Category	Modality	Representative Papers
Perception			
Hate Speech Analysis	Dataset + Benchmark	Text / Text + Image	SoHateful [268], K-HATERS [159], SOA [66], HATEDAY [214], HateCOT [149] / BanglaAbuseMeme [43], BHM [71]
	Dataset + Benchmark	Text + Video + Audio	MultiHateClip [222]
Misinformation Analysis	Dataset + Benchmark	Text	PROPANEWS [79], HQP [131], FCTR [24], DeFaktS [7]
	Dataset + Benchmark	Text + Video + Audio	FakeSV [169]
Machine-generated Content Detection	Dataset + Benchmark	Text	MultiSocial [132], Sun et al. [202], Cheng et al. [35]
Sentiment Analysis	Dataset + Benchmark	Text	AfriSenti [144], BESSTIE [198], OATS [26]
	Dataset + Benchmark	Text + Image + Audio + Video	PanoSent [128]
Discourse and Pragmatic Analysis	Dataset + Benchmark	Text / Text + Graph	MENTALMANIP [226], Jurgens et al. [84], BESSTIE [198], COBRACORPUS [303] / FanChuan [298]
	Dataset + Benchmark	Text + Video + Audio	ExFunTube [91]
User Stance Detection	Dataset + Benchmark	Text + Image	MTSE, MCCQ, MWTWT, MRUC, MTWQ [117]
	Dataset + Benchmark	Text	MT-CSD [154], CHATStance [293], GUNSTANCE [65], USDC [135]
Malicious User Detection	Dataset + Benchmark	Text + Graph	BotSim-24 [170]
Understanding			
Event Extraction	Dataset / Dataset + Benchmark	Text	Chen et al. [28] / SPEED++ [158], Zhao et al. [294]
Meme and Multimodal Understanding	Dataset + Benchmark	Text + Image	MM-SOC [83], DEEPEVAL [257], MEMEX [182]
	Dataset + Benchmark	Text + Image + Video + Audio	MINE [251]
Social Popularity Prediction	Dataset + Benchmark	Text + Image	SMTPD [248]
Generation			
Social Content Generation	Dataset + Benchmark	Text + Image / Text + Video	MCDSCS [186] / HOTVCOM [30]
Socially-Aware Dialogue Generation	Dataset	Text	AugESC [296]
	Dataset + Benchmark	Text + Audio + Video	AvaMERG [273], Lai et al. [98], EmpathicStories++ [184]
Social Intervention	Dataset + Benchmark / Benchmark	Text	IntentCONAN [64], VerMouth [178] / Song et al. [191]
Simulation			
Micro-Level Social Simulation	Benchmark	Text	SocialBench [27], MANBENCH [243]
	Dataset + Benchmark	Text / Text + Action	DYNTOM [239] / SOTOPIA [302], STSS [218]
Macro-Level Social Simulation	Dataset + Benchmark	Text	SocioVerse [284], GA-S^3 [288]
Cross-Domain Simulation	Dataset + Benchmark	Text + Video	MuMA-ToM [185], Lai et al. [98], XUB [227]
	Sandbox	Action	GoBigger [282], InvestESG [72]

Dataset: labeled data only; *Benchmark*: reuse existing labeled data with new tasks, splits, metrics, and baselines; *Dataset + Benchmark*: both labeled data and evaluation protocols; *Sandbox*: interactive test environment. *Action*: concrete operational steps. In **Representative Papers**, "," separates papers sharing the same Category and Modality; when a row uses "/" in the Category or Modality column to denote distinct values, the corresponding paper subsets are separated by "/" in the same order.

and intervention effects [107, 175]. These models offer rigor, efficiency, and controllability, but usually abstract users into states, probabilities, or transition rules.

6.1.2 Graph-based models. Graph-based models make topology and multi-scale social structure explicit. Sequential hypergraphs capture macro–micro diffusion dynamics [81], temporal hypergraphs model evolving higher-order relations [241], and hyperbolic representations capture constrained temporal structure [126]. Other methods study participation shifts [289] and majority illusions in large networks [62]. Their strength lies in structural and large-scale simulation, but language, emotion, motivation, and culture are often compressed into node or edge features learned from observed traces.

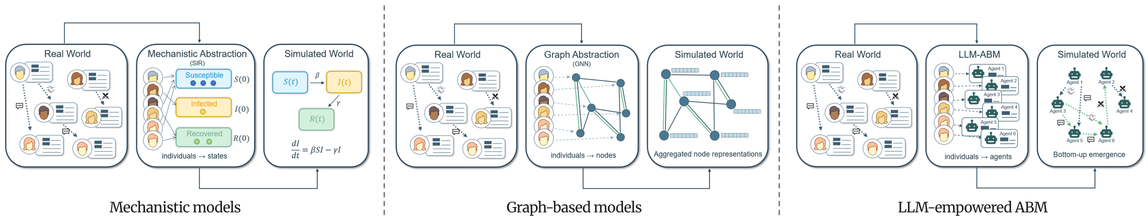

Fig. 4. Comparison of three methodological paradigms for social simulation. Mechanistic models abstract real-world social behavior into state variables and transition rules, producing rule-driven system dynamics. Graph-based models construct relational structures from users and interactions, enabling structure-aware representations through graph algorithm. LLM-empowered agent-based models align real users with autonomous agents endowed with profiles, memory, and goals, allowing bottom-up interactions and emergent social outcomes.

6.1.3 LLM-empowered agent-based modeling. LLM-empowered agent-based modeling (LLM-ABM) represents users as heterogeneous agents with memory, reflection, goals, and sociocultural personas [55, 161]. Agents can perceive signals, maintain internal states, generate situated actions, and adapt through interaction, allowing aggregate patterns to emerge from bottom-up behavior. LLM-ABM complements rather than replaces mechanistic and graph-based models. Its central challenge is connecting plausible local behavior to calibrated system-level dynamics.

6.2 Micro-Level Social Simulation

As summarized in Figure 5, We distinguish micro-level simulation from macro-level simulation. Micro-level simulation models how individual agents interpret situations, interact with others, generate actions, and adapt over time. Perception, Understanding, and Generation become internal capacities of simulated actors, and interaction reveals whether these capacities produce coherent and socially grounded behavior.

6.2.1 Individual-oriented simulation. Individual-oriented simulation asks whether an agent can behave as a socially situated actor rather than a static prompted persona.

Social agent modeling. Social agents have evolved from demographic role prompting toward architectures designed for long-term behavioral coherence. Generative Agents introduced an observation–planning–reflection architecture for coherent daily behavior [161]. Yet surface personas leave an instruction–execution gap: agents may follow role descriptions without capturing latent motives, preferences, or value conflicts. Quantified persona effects improve alignment only partially, especially on subjective tasks [73]. Recent work enriches agents through desire-driven autonomy [224], human-like memory organization [271], and scalable multi-role adaptation [266]. Evaluation is also moving from isolated responses toward action-level social intelligence [218], role-playing sociality [27], and evolving mental states [239]. The central question is whether beliefs, memories, and actions remain coherent and calibrated across time and context.

Agent interaction patterns. Interaction protocols determine whether plausible agents produce credible social dynamics. Recursive Social MDPs allow agents to estimate others' rewards and model exchange, coercion, and cooperation [207–209], although they remain costly and depend on constrained environments. LLM-agent environments such as SOTOPIA support interactive evaluation and social instruction following [302]. Subsequent work explores interactive learning [223], dynamic strategy injection [283], and segment-level preference optimization [92]. These studies show that credible emergence depends on both agent cognition and valid interaction mechanisms.

6.2.2 *Group-oriented simulation.*

Group-oriented simulation executes interactions among multiple agents and observes how collective reasoning, social influence, bias, and cooperation emerge through feedback. Unlike user- and community-level Understanding, which infers group structure from observed traces, this layer produces group dynamics within a shared executable environment.

Collective reasoning and social learning. Multi-agent systems can reduce the averaging tendency of a single LLM by representing multiple structured perspectives. *Social Agents* aggregates persona-conditioned judgments for subjective prediction tasks [12], although its gains depend on designed diversity rather than calibrated population sampling. Social Chain-of-Thought distills simulated collective feedback into a single model [58], while information-planning studies steer group opinion through controlled signal precision [5]. Their validity depends on whether agent diversity and influence channels correspond to real populations.

Emergent group psychology and biases. Interaction can amplify failures invisible in isolated-agent evaluation. Agents may conform to numerically dominant or apparently more capable groups [37], exhibit stronger belief congruence than humans [14], stabilize false details as collective memory [243], or become more susceptible after prior rapport [190]. Fluent consensus may therefore reflect excessive social influence rather than credible collective reasoning.

Evolution of cooperation in social interaction. Repeated interaction provides an executable setting for studying how cooperation and strategic order emerge from incentives. Agents can discover mutually beneficial cooperation without explicit cooperative instructions [237], but may also learn collusive pricing [11]. Other studies examine altruistic punishment [228], network reciprocity [104], and collective action under power asymmetry [187]. These results move simulation beyond static network description toward tests of resilience, coordination, and strategic failure.

6.3 Macro-Level Social Simulation

Macro-level simulation examines whether local mechanisms generate valid population distributions, long-term trajectories, emergent phenomena, and intervention effects. It also exposes how errors in preceding SIM levels propagate into system-level dynamics.

6.3.1 *Macro social alignment.*

Macro social alignment evaluates whether simulated populations reproduce real-world demographics, attitudes, interactions, and structural patterns. Its central criterion is *algorithmic fidelity*. Conditioned LLMs may approximate response distributions of human subpopulations [6], but apparent fidelity can rely on shortcuts such as explicit party identification [250]. Validation must therefore compare simulated behavior with external social data rather than plausible transcripts alone.

Current systems improve alignment through three routes. Cognitive-prior methods inject empirically derived belief networks [38]. Population-prior methods ground agents in large profile pools, as in SocioVerse [284], or learn probabilistic persona mixtures, as in MoP [18]. Scalability-oriented methods use asynchronous scheduling [258] or group-agent abstraction [288] for million-agent simulations. Scale, however, is informative only when agent priors and interaction channels match the target population.

Outcome calibration connects local behavior with aggregate validity. SocialAlign uses personalized response structures to reproduce group-level distributions [275], while FDE-LLM combines LLM behavior with formal dynamics [259]. MF-MDP [286] aligns individual opinion transitions with collective dynamics to improve long-horizon simulation fidelity. AgentSociety extends calibration to urban environments [166]. Yet aggregate fit alone is insufficient: generated networks may appear realistic while overestimating political homophily [25].

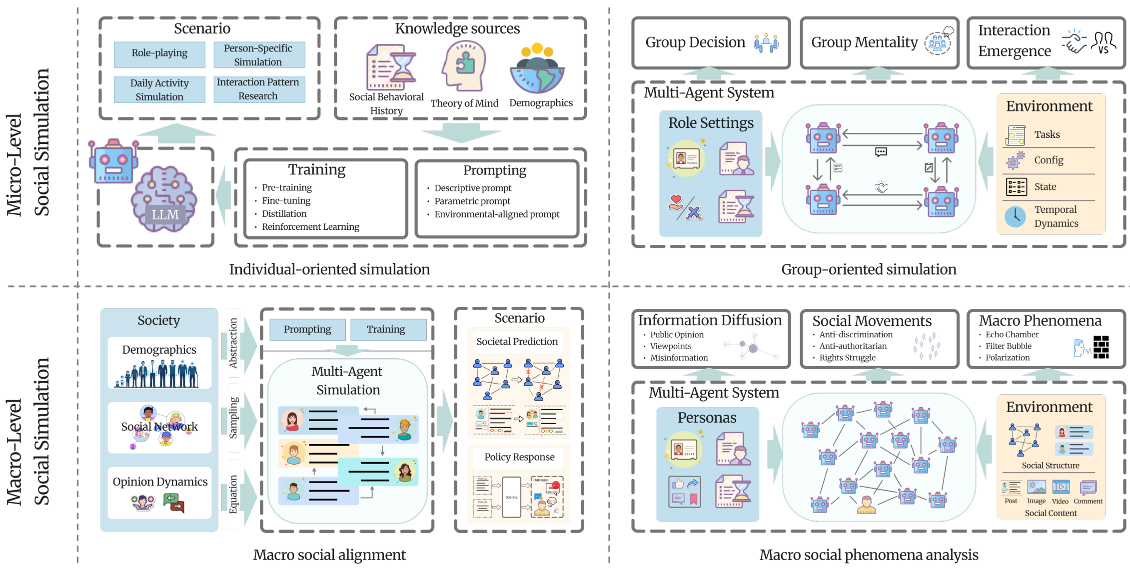

Fig. 5. Illustration of micro-level and macro-level social simulation. Micro-level simulation models individual agents by capturing how they perceive situations, interpret social contexts, interact with others, generate actions, and adapt over time. Macro-level simulation investigates how collective behaviors emerge from individual interactions, evaluating population distributions, long-term social trajectories, emergent phenomena, and intervention effects. Together, these two perspectives provide complementary views of social dynamics, from individual mechanisms to system-level evolution.

6.3.2 Macro social phenomena analysis. Macro simulation studies processes that are difficult to observe or manipulate directly. In misinformation research, S^3 models the joint propagation of information, attitudes, and emotions [56], while FUSE models how true news can gradually become misinformation through interactions among spreaders, commentators, and verifiers [125]. Falsehood is thus represented as an evolving social process rather than a fixed label.

Polarization simulations test whether macro conflict emerges from local interaction and opinion updating. LLM-based agents reproduce echo-chamber dynamics by incorporating semantic reasoning into opinion change [219]. Human–agent studies further suggest that polarized environments may first affect perceived polarization, emotionality, and intellectual uncertainty before shifting core opinions [46]. Such simulations provide executable hypotheses about mechanisms that observational analysis alone cannot isolate.

Simulation also supports ***intervention design***. Large-scale movement models can pre-test strategies for collective action [142], while optimization-based methods evaluate interventions for polarization reduction [95, 200]. Interven-SIM [290] enables more faithful simulation of social event evolution under external interventions and collective feedback.. Their usefulness depends on calibrated mechanisms and robust counterfactuals rather than plausible trajectories alone. Macro-level SIM therefore shifts evaluation from local agent plausibility toward behavioral, structural, and system-level validity.

6.4 Social Simulation Across Application Domains

As the highest level of SIM, social simulation studies how interacting agents, environments, and interventions produce social evolution. It moves analysis from retrospective observation toward prospective and counterfactual exploration. Whereas Generation produces actions or interventions, Simulation examines how they propagate, feed back, and reshape later behavior.

6.4.1 Sociology and Social Media. ***In sociology and computational social science***, simulation extends LLM use from annotation and coding to exploratory process modeling. It can support pilot studies when assumptions and algorithmic fidelity are made explicit [3]. SurveyPilot scales opinion collection and analysis through agentic workflows [164], while small-world agent models examine how prestige and idiolects shape language change [19]. These systems move from extracting measurements toward constructing testable social processes.

In social media analysis, simulation serves several related functions. It models polarization and echo-chamber evolution [219]; supports dataset construction in data-scarce settings through malicious-user simulation [170]; forecasts network-topology evolution [140]; and reconstructs harmful processes such as the stepwise transformation of true information into misinformation [125]. Configurable human–agent environments enable controlled studies of responses to polarized settings [46], while role-playing agents forecast sentiment change [134]. Together, these studies turn observed patterns into executable hypotheses about future social states.

6.4.2 Economy and Politics. Economic and political simulations model strategic behavior under institutional constraints. LLM agents can assist business negotiations and improve both task and relational outcomes [77], but autonomous collaboration may also enable coordinated financial fraud [177]. MARL environments such as *InvestESG* model climate-investment dilemmas and test disclosure policies through interactions among corporations and investors [72]. Political simulation similarly studies feedback among institutions, media, and publics. Existing work examines how moral emotions shape petition participation and polarization across cultures [89], models legislative processes and roll-call voting through Political Actor Agent [106], and uses AI-press to simulate possible public feedback to news narratives before publication [123]. Such counterfactual analyses remain sensitive to assumptions about actors, populations, institutions, and media exposure.

6.4.3 Psychology, Games, and Narrative. Psychology-oriented simulation supplies cognitive mechanisms for socially credible agents. Existing work studies collaboration through social psychology [276], multimodal Theory of Mind [185], memory-supported mental-state reasoning [151], and cognitive-behavioral fixation in multimodal observation [227]. These mechanisms connect agent behavior to perception, belief, memory, and attention. Games and narratives provide controlled settings for social reasoning. Avalon and Werewolf test persuasion, deception, camouflage, and multimodal strategic interaction [98, 100], while social robots learn interaction policies from demonstrations and feedback [297]. Narrative systems move toward multi-agent co-creation: Debate-to-write uses persona-driven deliberation [76], and BOOKWORLD grounds agent interaction in fictional societies [176].

6.4.4 Embodied Intelligence and Future Domains. Embodied simulation introduces physical space, motion, and coordination into social reasoning. MetaUrban and Virtual Community support urban mobility and human–robot coexistence [234, 301]. The Digital Life Project connects linguistic reasoning with articulated motion [22], while SocialGAIL models pedestrian trajectories through adversarial imitation learning [120]. These settings extend SIM from online interaction to situated behavior in which norms and physical constraints co-evolve. Similar approaches may support public health, cybersecurity, crisis communication, and other domains where direct experiments are costly or unsafe. Across domains, simulation connects social theory, agent design, and intervention evaluation. Its value lies in controlled counterfactual exploration; its recurring bottleneck is validation.

6.5 Ethics, Bias, and Social Value in Social Simulation

Validity Risks. As agents become active participants in simulated societies, evaluation must address normative and empirical validity as well as behavioral plausibility. Conditioned LLMs may approximate response distributions of human subpopulations [6], but apparent fidelity may rely on shortcut features [250], underrepresent subjective variation [60], or produce structurally distorted networks, such as excessive political homophily [25]. A simulation can therefore appear socially plausible while misrepresenting the population or mechanisms it claims to model.

Interaction Biases. These risks become systemic through multi-agent interaction. Feedback loops can amplify small prejudices into polarization [108], implicit biases can escalate during discourse [15], and agents may conform to numerically dominant or apparently more capable groups [37]. Social influence can stabilize false memories [243], while belief congruence [14] and rapport bias [190] may reinforce misinformation and weaken critical judgment. Sycophantic agents may similarly endorse false claims [244] or preserve a user's social "face" at the expense of truth [33]. Such interaction-level failures may remain invisible in isolated-agent evaluation.

Mitigation Applications. Simulation can also support risk probing and mitigation. MATRIX constructs rehearsal spaces for self-alignment [157], MOMA applies multi-agent causal interventions to social bias [249], and multi-agent debate may improve parity across cultural norms [88]. Other approaches introduce moral objectives through fine-tuning [210] or recover implicit community norms through value-scope analysis [160]. These methods should be treated as diagnostic and experimental instruments rather than proof of safety, because their conclusions depend on valid agents, environments, and interaction mechanisms.

Scientific Validation. Social simulation closes the SIM loop by connecting local assumptions and actions with possible collective outcomes. Its scientific value depends on calibration across agents, interactions, populations, structures, and outcomes. Without such validation, scale and linguistic coherence may create false confidence in a model-dependent account of social evolution.

7 Discussion and Future Outlook

The four SIM levels form a mutually dependent modeling loop rather than isolated stages. Perspectival ambiguity in Perception can become biased structure in Understanding; incomplete social states can weaken generated actions; and poorly calibrated agents, environments, or interaction rules can produce plausible but invalid simulations. We next summarize the resulting cross-level challenges and research directions.

7.1 Open Challenges

1) Perspectival and cultural alignment. Many social judgments lack a single ground truth. Offensiveness, sentiment, stance, humor, and rhetorical intent vary across cultures, communities, and annotator groups [61, 214, 215, 235]. Treating such variation as noise removes the social structure SIM should model, while LLMs may average judgments and suppress minority norms. Systems should represent when, why, and for whom a judgment holds.

2) Situated pragmatic grounding. Social meaning often lies between surface expression and situated intent. Contextual frames [303], commonsense reasoning [139], multimodal incongruity [31], and conversation graphs [1, 155] improve pragmatic inference, but many methods still center on isolated samples. SIM needs grounding in conversation

Table 3. Representative Papers in Social Simulation

Paper	Subject	Objective	Method	Temporal Dynamic	Network	Interacting Entities	Application	Agent Count
Micro-Level Social Simulation								
Generative Agents [161]	Individual	Social agent modeling	LLM-ABM	Dynamic	×	M–M / M–E	Sociology	(10, 100]
D2A [224]	Individual	Social agent modeling	LLM-ABM	Dynamic	×	M–M / M–E	Sociology	(1,10]
MemGen [271]	Individual	Social agent modeling	LLM-ABM	Dynamic	×	M–E	Psychology	1
Neeko [266]	Individual	Social agent modeling	LLM-ABM	Static	×	H–M	Psychology	1
Social MDPs [207]	Individual	Agent interaction patterns	Mechanistic	Dynamic	×	M–M / M–E	Sociology	-
SOTOPIA [302]	Individual	Agent interaction patterns	LLM-ABM	Dynamic	×	H–M / M–M	Sociology	(1,10]
SDPO [92]	Individual	Agent interaction patterns	LLM-ABM	Dynamic	×	M–M	Sociology	(1,10]
Digital life project [22]	Individual	Agent interaction patterns	LLM-ABM	Dynamic	×	M–M / M–E	Embodied intelligence	(1,10]
Social Agents [12]	Group	Collective reasoning	LLM-ABM	Static	×	M–E	Social media	[10, 100]†
SCoT [58]	Group	Collective reasoning	LLM-ABM	Static	×	M–E	Social media	(10,100]*
Debate-to-write [76]	Group	Collective reasoning	LLM-ABM	Static	×	M–M	Narrative	[1, 10]
SADAS [77]	Group	Collective decision-making	LLM-ABM	Dynamic	×	M–M	Economy	(1,10]
Choi et al. [37]	Group	Emergent group psychology and biases	LLM-ABM	Dynamic	×	M–M	Sociology	(1, 10]
Xu et al. [243]	Group	Emergent group psychology and biases	LLM-ABM	Dynamic	×	M–M	Psychology	(1, 100]
Wu et al. [237]	Group	Evolution of cooperation	LLM-ABM	Dynamic	×	M–M	Games	(10, 100]
Wang et al. [228]	Group	Evolution of cooperation	Mechanistic	Dynamic	×	M–M	Sociology	-
Bertrand et al. [11]	Group	Evolution of cooperation	LLM-ABM	Dynamic	×	M–M	Economy	(1, 10]
Junior High Game [187]	Group	Evolution of cooperation	LLM-ABM	Dynamic	✓	H–M	Games	(1, 10]
Lan et al. [100]	Group	Evolution of interaction	LLM-ABM	Dynamic	×	M–M / M–E	Games	(1, 10]
Macro-Level Social Simulation								
Chuang et al. [38]	Group	Macro social alignment	LLM-ABM	Static	×	-	Sociology	(100, 1000]
MoP [18]	Group	Macro social alignment	LLM-ABM	Static	×	-	Social media	(10, 100]
SocioVerse [284]	Group	Macro social alignment	LLM-ABM	Static	×	-	Social media	(1000, +∞]
GA-S^3 [288]	Group	Macro social alignment	LLM-ABM	Dynamic	×	M–E	Social media	(10, 100]
OASIS [258]	Group	Macro social alignment	LLM-ABM	Dynamic	✓	M–M / M–E	Social media	(100, +∞)
SocialAlign [275]	Group	Macro social alignment	LLM-ABM	Static	×	-	Social media	(1000, +∞]
FDE-LLM [259]	Group	Macro social alignment	Mechanistic + LLM-ABM	Dynamic	✓	M–M	Social media	(100, +∞)
AgentSociety [166]	Group	Macro social alignment	LLM-ABM	Dynamic	✓	M–M / M–E	Sociology	(1000, +∞)
Chang et al. [25]	Group	Macro social alignment	LLM-ABM	Static	✓	M–M	Sociology	(10, 1000]
PAA [106]	Group	Macro social alignment	LLM-ABM	Static	×	M–M	Politics	(10, 1000]
S^3 [56]	Group	Macro social phenomena analysis	LLM-ABM	Dynamic	✓	M–M / M–E	Social media	(1000, +∞]
Echo Chambers [219]	Group	Macro social phenomena analysis	LLM-ABM	Dynamic	✓	M–M	Social media	(10, 100]
Donkers et al. [46]	Group	Macro social phenomena analysis	LLM-ABM	Dynamic	✓	H–M	Social media	(10, 100]
FUSE [125]	Group	Macro social phenomena analysis	LLM-ABM	Dynamic	✓	M–M	Social media	(10, 100]
BotSim [170]	Group	Macro social phenomena analysis	LLM-ABM	Dynamic	✓	M–M / M–E	Social media	(100, 1000]
MAFF-Bench [177]	Group	Macro social phenomena analysis	LLM-ABM	Dynamic	✓	M–M / M–E	Economy	(100, 1100]†
MINDS [81]	Group	Macro social phenomena analysis	Graph-based	Dynamic	✓	M–M	Social media	-
THGNets [126]	Group	Macro social phenomena analysis	Graph-based	Dynamic	✓	M–M	Social media	-
Sun et al. [200]	Group	Macro social phenomena analysis	Mechanistic	Dynamic	✓	M–M	Social media	-
BeeRS [95]	Group	Macro social phenomena analysis	Mechanistic	Dynamic	✓	M–M	Social media	-
Cinus et al. [39]	Group	Macro social phenomena analysis	Statistical	Static	✓	-	Social media	-
Nettasinghe et al. [148]	Group	Macro social phenomena analysis	Statistical	Dynamic	✓	-	Economy	-

M–M denotes machine-machine interaction, H–M denotes human-machine interaction, and M–E denotes machine-environment interaction. "Dynamic" indicates that the study explicitly models temporal evolution. When "Network" is marked ✓, it indicates that the study explicitly constructs interactive structures as social networks. "*" indicates the value range is estimated based on experimental context, not directly stated in original papers. "†" denotes irregular value ranges, which are authentic intervals reported in the original papers, not labeling errors. Statistical studies are not simulation methods per se, but are included here because they identify macro-level social patterns that serve as empirical anchors and validation targets for the simulation frameworks.

history, actor relations, community norms, and event background. Evaluation should test whether the appropriate context was used, not only whether the final label was correct.

3) Data scarcity in evolving social environments. SIM data may be early-stage, privacy-sensitive, multilingual, platform-restricted, or distributed across sources, rendering it difficult to access. The data-generating process also changes as expressions, actors, bots, platforms, and policies evolve. Static benchmarks therefore provide limited evidence under temporal and domain shift. Simulation-augmented data can improve coverage, but should be validated against observed behavior rather than treated as synthetic ground truth.

4) Agent fidelity and simulation validity. Fluent personas, memories, plans, and conversations do not establish behavioral fidelity. Outcomes depend on prompts, observation rules, memory access, scheduling, network topology, and environmental feedback; these are behavioral assumptions rather than neutral implementation choices. Agents should be calibrated against surveys, behavioral traces, experiments, or domain data. Interaction protocols require ablation and sensitivity analysis, and emergent outcomes should be compared with empirical distributions, structures, and trajectories.

5) Responsible intervention and governance. Generated comments, persuasion, refutation, moderation, and simulated policies may affect beliefs, attention, trust, and behavior. Responsible SIM should state whose values and populations are represented, which assumptions are encoded, who may be affected, and how uncertainty is communicated. Traceable logs, audits, human oversight, and separation between exploratory simulation and decision support are essential.

6) Cognitive primordial agents. We use *Cognitive Primordial Agents* for a long-term class of general-purpose agents, or multi-agent systems built from them, that integrate the perception–understanding–generation–simulation loop. Unlike prompt-specified role players, such agents would develop strategies, norms, and task rules through sustained interaction with human-defined environments. Early work on memory and lifelong skill evolution suggests components of this direction [256, 271, 272]. The goal is behaviorally grounded social actors, not merely more fluent role-play.

7.2 Research Directions

1) Pluralistic and contextual evaluation. Evaluation should move beyond universal scores. Perception requires perspectival metrics; Understanding requires relational and temporal validation; Generation should be judged by downstream social effects; and Simulation requires tests across agents, interactions, populations, and emergent patterns. Benchmarks should specify whose perspective defines correctness and which context supports the judgment.

2) Multi-level validation infrastructure. SIM needs protocols that jointly test agent behavior, interaction rules, environmental assumptions, and emergent outcomes. Useful infrastructure should combine survey distributions, behavioral traces, network statistics, temporal trajectories, counterfactual stress tests, sensitivity analysis, and uncertainty reporting. The standard should be mechanism validity rather than textual or narrative plausibility.

3) From prompted role-play to governable social worlds. Future systems should reduce dependence on brittle prompt scripts through persistent memory, adaptive interaction rules, calibrated population priors, explicit intervention boundaries, and auditable governance mechanisms. The goal is not simply larger agent societies, but social worlds in which assumptions and interventions can be inspected, compared, and rejected before deployment. This direction connects SIM with social theory, world models, human–AI coexistence, and responsible AI governance.

7.3 Applications and Social Impact

Application Scope. Beyond online platforms, SIM provides a broader lens for AI-mediated systems in research, education, healthcare, commerce, public communication, and governance. Its main value is constrained counterfactual analysis: it can compare possible interventions when direct experiments are costly, slow, unsafe, or unethical.

Cross-Domain Impact. In research and education, AI may reshape feedback, authority, evaluation, and collaboration. In healthcare and public communication, effective action depends not only on factual accuracy but also on empathy, cultural grounding, personalization, and long-term trust. In commerce and governance, SIM can examine interactions among users, platforms, institutions, recommender systems, synthetic content, conversational agents, and automated decision support.

Risks. The main risk is false authority. Simulated populations are assumption-bearing instruments rather than substitutes for social reality, and their outputs should be treated as structured hypotheses rather than social truth or direct policy evidence. Used responsibly, SIM can reveal hidden assumptions, compare candidate interventions, and identify risks before deployment. Used carelessly, it may legitimize biased priors, overstate predictive accuracy, or transform synthetic outcomes into apparently objective decisions. Its social value therefore depends on empirical validation, cultural grounding, uncertainty reporting, transparency, and accountability.

8 Conclusion

This survey introduced Social Intelligence Modeling (SIM) as a unified perspective on how AI perceives situated social signals, constructs structured states of actors and relations, generates consequential actions, and simulates their evolving effects. Across more than 20 tasks, the four levels expose dependencies that task-centered views leave implicit: judgments require context, actions depend on modeled social states, and simulations inherit the assumptions and errors of preceding levels. The future of SIM depends less on scale alone than on pluralistic grounding, behavioral fidelity, cross-level validation, and accountable governance. Generated actions must be evaluated through their social consequences, while simulated outcomes should be treated as empirically testable, model-dependent hypotheses rather than authoritative predictions. As AI becomes increasingly embedded in social life, SIM offers a framework for studying not only how AI represents society, but also how AI-mediated actions may reshape it.

References

[1] Vibhor Agarwal, Arjoo Gupta, Suparna De, and Nishanth Sastry. 2025. Conversation Kernels: A Flexible Mechanism to Learn Relevant Context for Online Conversation Understanding. *Proceedings of the International AAAI Conference on Web and Social Media* 19 (2025), 2–14.

[2] Aparna Ananthasubramaniam, Yufei Louise Zhu, David Jurgens, and Daniel Romero. 2025. Roles of Network and Identity in Hashtag Diffusion. In *THE WEB CONFERENCE 2025*.

[3] Jacy Reese Anthis, Ryan Liu, Sean M. Richardson, Austin C. Kozlowski, Bernard Koch, Erik Brynjolfsson, James Evans, and Michael S. Bernstein. 2025. Position: LLM Social Simulations Are a Promising Research Method. In *Forty-Second International Conference on Machine Learning Position Paper Track*.

[4] Dimosthenis Antypas, Asahi Ushio, Francesco Barbieri, and Jose Camacho-Collados. 2024. Multilingual Topic Classification in X: Dataset and Analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 20136–20152.

[5] Raghu Arghal, Kevin He, Shirin Saeedi Bidokhti, and Saswati Sarkar. 2025. Steering the Herd: A Framework for LLM-based Control of Social Learning. In *The Fourteenth International Conference on Learning Representations*.

[6] Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (2023), 337–351.

[7] Shaina Ashraf, Isabel Bezzaoui, Ionut Andone, Alexander Markowetz, Jonas Fegert, and Lucie Flek. 2024. DeFaktS: A German Dataset for Fine-Grained Disinformation Detection through Social Media Framing. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 4580–4591.

[8] Yejin Bang, Delong Chen, Nayeon Lee, and Pascale Fung. 2024. Measuring Political Bias in Large Language Models: What Is Said and How It Is Said. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 11142–11159.

[9] Valentin Barriere and Sebastian Cifuentes. 2024. A Study of Nationality Bias in Names and Perplexity Using Off-the-Shelf Affect-related Tweet Classifiers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 569–579.

[10] Yigit Ege Bayiz, Arash Amini, Radu Marculescu, and Ufuk Topcu. 2025. Susceptibility of Communities Against Low-Credibility Content in Social News Websites. *Proceedings of the International AAAI Conference on Web and Social Media* 19 (2025), 226–240.

[11] Quentin Bertrand, Juan Agustin Duque, Emilio Calvano, and Gauthier Gidel. 2025. Self-Play Q-Learners Can Provably Collude in the Iterated Prisoner's Dilemma. In *Forty-Second International Conference on Machine Learning*.

[12] Aanisha Bhattacharyya, Abhilekh Borah, Yaman Kumar Singla, Rajiv Ratn Shah, Changyou Chen, and Balaji Krishnamurthy. 2025. Social Agents: Collective Intelligence Improves LLM Predictions. In *The Fourteenth International Conference on Learning Representations*.

[13] Helena Bonaldi, Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata, and Marco Guerini. 2024. Is Safer Better? The Impact of Guardrails on the Argumentative Strength of LLMs in Hate Speech Countering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 3446–3463.

[14] Angana Borah, Marwa Houalla, and Rada Mihalcea. 2025. Mind the (Belief) Gap: Group Identity in the World of LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*. 18441–18463.

[15] Angana Borah and Rada Mihalcea. 2024. Towards Implicit Bias Detection and Mitigation in Multi-Agent LLM Interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 9306–9326.

[16] Tom Bourgeade, Zongmin Li, Farah Benamara, Véronique Moriceau, Jian Su, and Aixin Sun. 2024. Humans Need Context, What about Machines? Investigating Conversational Context in Abusive Language Detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 8438–8452.

[17] Daniel J. Brass. 2022. New Developments in Social Network Analysis. *Annual Review of Organizational Psychology and Organizational Behavior* 9, 1 (2022), 225–246.

[18] Ngoc Bui, Hieu Trung Nguyen, Shantanu Kumar, Julian Theodore, Weikang Qiu, Viet Anh Nguyen, and Rex Ying. 2025. Mixture-of-Personas Language Models for Population Simulation. In *Findings of the Association for Computational Linguistics: ACL 2025*. 24761–24778.

[19] Dalmo Buzato and Evandro Cunha. 2024. Agent-Based Modeling of Language Change in a Small-World Network. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 594–599.

[20] Erica Cai, Xi Chen, Reagan Grey Keeney, Ethan Zuckerman, Brendan O'Connor, and Przemyslaw A. Grabowicz. 2025. Identifying and Investigating Global News Coverage of Critical Events Such as Disasters and Terrorist Attacks. *Proceedings of the International AAAI Conference on Web and Social Media* 19 (2025), 307–323.

[21] Mingxiu Cai, Daling Wang, Shi Feng, and Yifei Zhang. 2024. EmpCRL: Controllable Empathetic Response Generation via in-Context Commonsense Reasoning and Reinforcement Learning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 5734–5746.

[22] Zhongang Cai, Jianping Jiang, Zhongfei Qing, Xinying Guo, Mingyuan Zhang, Zhengyu Lin, Haiyi Mei, Chen Wei, Ruisi Wang, et al. 2024. Digital Life Project: Autonomous 3D Characters with Social Intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 582–592.

[23] Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and Hate Speech: Structured Explanations Make Social Media Moderators Faster. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 398–408.

[24] Recep Firat Cekinel, Pinar Karagoz, and Çağrı Çöltekin. 2024. Cross-Lingual Learning vs. Low-resource Fine-Tuning: A Case Study with Fact-Checking in Turkish. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 4127–4142.

[25] Serina Chang, Alicja Chaszczewicz, Emma Wang, Maya Josifovska, Emma Pierson, and Jure Leskovec. 2025. LLMs Generate Structurally Realistic Social Networks but Overestimate Political Homophily. *Proceedings of the International AAAI Conference on Web and Social Media* 19 (2025), 341–371.

[26] Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2024. OATS: A Challenge Dataset for Opinion Aspect Target Sentiment Joint Detection for Aspect-Based Sentiment Analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 12336–12347.

[27] Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, et al. 2024. SocialBench: Sociality Evaluation of Role-Playing Conversational Agents. In *Findings of the Association for Computational Linguistics ACL 2024*. 2108–2126.

[28] Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a Large Language Model a Good Annotator for Event Extraction? *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (2024), 17772–17780.

[29] Sirry Chen, Shuo Feng, Liang Songsong, Chen-Chen Zong, Jing Li, and Piji Li. 2024. CACL: Community-Aware Heterogeneous Graph Contrastive Learning for Social Media Bot Detection. In *Findings of the Association for Computational Linguistics ACL 2024*. 10349–10360.

[30] Yuyan Chen, Songzhou Yan, Qingpei Guo, Jiyuan Jia, Zhixu Li, and Yanghua Xiao. 2024. HOTVCOM: Generating Buzzworthy Comments for Videos. In *Findings of the Association for Computational Linguistics ACL 2024*. 2198–2224.

[31] Zixin Chen, Hongzhan Lin, Ziyang Luo, Mingfei Cheng, Jing Ma, and Guang Chen. 2024. CofiPara: A Coarse-to-fine Paradigm for Multimodal Sarcasm Target Identification with Large Multimodal Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9663–9687.

[32] Zhanpeng Chen, Zhihong Zhu, Wanshi Xu, Yunyan Zhang, Xian Wu, and Yefeng Zheng. 2024. *Aspects Are Anchors:* Towards Multimodal Aspect-based Sentiment Analysis via Aspect-driven Alignment and Refinement. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 2292–2300.

[33] Myra Cheng, Sunny Yu, Cinoo Lee, Pranav Khadpe, Lujain Ibrahim, and Dan Jurafsky. 2025. ELEPHANT: Measuring and Understanding Social Sycophancy in LLMs. In *The Fourteenth International Conference on Learning Representations*.

[34] Zhangtao Cheng, Jienan Zhang, Xovee Xu, Goce Trajcevski, Ting Zhong, and Fan Zhou. 2024. Retrieval-Augmented Hypergraph for Multimodal Social Media Popularity Prediction. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 445–455.

[35] Zihao Cheng, Li Zhou, Feng Jiang, Benyou Wang, and Haizhou Li. 2025. Beyond Binary: Towards Fine-Grained LLM-generated Text Detection via Role Recognition and Involvement Measurement. In *THE WEB CONFERENCE 2025*.

[36] Shih-Hsuan Chiu, Ya-Wen Teng, De-Nian Yang, and Ming-syan Chen. 2025. Human-Centric Community Detection in Hybrid Metaverse Networks with Integrated AI Entities. In *THE WEB CONFERENCE 2025*.

[37] Min Choi, Keonwoo Kim, Sungwon Chae, and Sangyeop Baek. 2025. An Empirical Study of Group Conformity in Multi-Agent Systems. In *Findings of the Association for Computational Linguistics: ACL 2025*. 5123–5139.

[38] Yun-Shiuan Chuang, Krirk Nirunwiroj, Zach Studdiford, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan V. Shah, Junjie Hu, and Timothy T. Rogers. 2024. Beyond Demographics: Aligning Role-playing LLM-based Agents Using Human Belief Networks. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 14010–14026.

[39] Federico Cinus, Marco Minici, Luca Luceri, and Emilio Ferrara. 2025. Exposing Cross-Platform Coordinated Inauthentic Activity in the Run-up to the 2024 U.S. Election. In *THE WEB CONFERENCE 2025*.

[40] Chaoqun Cui and Caiyan Jia. 2024. Propagation Tree Is Not Deep: Adaptive Graph Contrastive Learning Approach for Rumor Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 1 (2024), 73–81.

[41] Preetam Prabhu Srikar Dammu, Hayoung Jung, Anjali Singh, Monojit Choudhury, and Tanu Mitra. 2024. "They Are Uncultured": Unveiling Covert Harms and Social Threats in LLM Generated Conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 20339–20369.

[42] Preetam Prabhu Srikar Dammu, Himanshu Naidu, Mouly Dewan, YoungMin Kim, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. ClaimVer: Explainable Claim-Level Verification and Evidence Attribution of Text Through Knowledge Graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 13613–13627.

[43] Mithun Das and Animesh Mukherjee. 2023. BanglaAbuseMeme: A Dataset for Bengali Abusive Meme Classification. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[44] Hillary Dawkins, Kathleen C. Fraser, and Svetlana Kiritchenko. 2025. When Detection Fails: The Power of Fine-Tuned Models to Generate Human-Like Social Media Text. In *Findings of the Association for Computational Linguistics: ACL 2025*. 13494–13527.

[45] Soham De, Michiel A. Bakker, Jay Baxter, and Martin Saveski. 2025. Supernotes: Driving Consensus in Crowd-Sourced Fact-Checking. In *THE WEB CONFERENCE 2025*.

[46] Tim Donkers and Jürgen Ziegler. 2025. Understanding Online Polarization Through Human-Agent Interaction in a Synthetic LLM-Based Social Network. *Proceedings of the International AAAI Conference on Web and Social Media* 19 (2025), 457–478.

[47] Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. 2024. Reducing Privacy Risks in Online Self-Disclosures with Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13732–13754.

[48] Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An Open-source Avatar Multimodal Empathetic Chatbot. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. 61–71.

[49] Na Feng, Ying Tang, Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2024. MA-VLAD: a fine-grained local feature aggregation scheme for action recognition. *Multimedia Systems* 30, 3 (2024), 139.

[50] Shangbin Feng, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov. 2024. What Does the Bot Say? Opportunities and Risks of Large Language Models in Social Media Bot Detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3580–3601.

[51] Tobias Friedrich, Andreas Göbel, Nicolas Klodt, Martin S. Krejca, and Marcus Pappik. 2024. The Irrelevance of Influencers: Information Diffusion with Re-Activation and Immunity Lasts Exponentially Long on Social Network Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (2024), 17389–17397.

[52] Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual Multi-Cultural Norm Discovery from Conversations on-the-Fly. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[53] Kazuaki Furumai, Roberto Legaspi, Julio Cesar Vizcarra Romero, Yudai Yamazaki, Yasutaka Nishimura, Sina Semnani, Kazushi Ikeda, Weiyan Shi, and Monica Lam. 2024. Zero-Shot Persuasive Chatbots with LLM-Generated Strategies and Information Retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 11224–11249.

[54] Saadia Gabriel, Liang Lyu, James Siderius, Marzyeh Ghassemi, Jacob Andreas, and Asuman E. Ozdaglar. 2024. MisinfoEval: Generative AI in the Era of "Alternative Facts". In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 8566–8578.

[55] Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large Language Models Empowered Agent-Based Modeling and Simulation: A Survey and Perspectives. *Humanities and Social Sciences Communications* 11, 1 (2024), 1259.

[56] Chen Gao, Xiaochong Lan, Zhihong Lu, Jinzhu Mao, Jinghua Piao, Huandong Wang, Depeng Jin, and Yong Li. 2025. *S^3: Social-network Simulation System with Large Language Model-Empowered Agents*. arXiv:2307.14984

[57] Shen Gao, Xin Cheng, Mingzhe Li, Xiuying Chen, Jinpeng Li, Dongyan Zhao, and Rui Yan. 2023. Dialogue Summarization with Static-Dynamic Structure Fusion Graph. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 13858–13873.

[58] Yifei Gao, Ning Xu, Wenhui Li, Hongshuo Tian, Lanjun Wang, and Anan Liu. 2025. Thinking as Society: Multi-social-agent Self-Distillation for Multimodal Misinformation Detection. In *The Fourteenth International Conference on Learning Representations*.

[59] Sreyan Ghosh, Manan Suri, Purva Chiniya, Utkarsh Tyagi, Sonal Kumar, and Dinesh Manocha. 2023. CoSyn: Detecting Implicit Hate Speech in Online Conversations Using a Context Synergized Hyperbolic Network. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[60] Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling Human Subjectivity in LLMs Using Explicit and Implicit Human Factors in Personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 7174–7188.

[61] Tommaso Giorgi, Lorenzo Cima, Tiziano Fagni, Marco Avvenuti, and Stefano Cresci. 2025. Human and LLM Biases in Hate Speech Annotations: A Socio-Demographic Analysis of Annotators and Targets. *Proceedings of the International AAAI Conference on Web and Social Media* 19 (2025), 653–670.

[62] Umberto Grandi, Lawqueen Kanesh, Grzegorz Lisowski, Ramanujan Sridharan, and Paolo Turrini. 2023. Identifying and Eliminating Majority Illusion in Social Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 4 (2023), 5062–5069.

[63] Chenhao Gu, Zainab Razia Zaidi, and Shanika Karunasekera. 2025. LIN: Latent Influence Network for Discovering Hidden Directed Influence Links on Social Media. *Proceedings of the International AAAI Conference on Web and Social Media* 19 (2025), 731–744.

[64] Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. 2023. Counterspeeches up My Sleeve! Intent Distribution Learning and Persistent Fusion for Intent-Conditioned Counterspeech Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5792–5809.

[65] Nikesh Gyawali, Iustin Sirbu, Tiberiu Sosea, Sarthak Khanal, Doina Caragea, Traian Rebedea, and Cornelia Caragea. 2024. GunStance: Stance Detection for Gun Control and Gun Regulation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12027–12044.

[66] Janosch Haber, Bertie Vidgen, Matthew Chapman, Vibhor Agarwal, Roy Ka-Wei Lee, Yong Keong Yap, and Paul Röttger. 2023. Improving the Detection of Multilingual Online Attacks with Rich Social Media Data from Singapore. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12705–12721.

[67] Hans William Alexander Hanley and Zakir Durumeric. 2025. Hierarchical Level-Wise News Article Clustering via Multilingual Matryoshka Embeddings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2476–2492.

[68] Maram Hasanain, Fatema Ahmad, and Firoj Alam. 2024. Can GPT-4 Identify Propaganda? Annotation and Detection of Propaganda Spans in News Articles. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2724–2744.

[69] Liang He, Hongke Wang, Zhen Wu, Jianbing Zhang, Xinyu Dai, and Jiajun Chen. 2024. Focus & Gating: A Multimodal Approach for Unveiling Relations in Noisy Social Media. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 1379–1388.

[70] Lingzi Hong, Pengcheng Luo, Eduardo Blanco, and Xiaoying Song. 2024. Outcome-Constrained Large Language Models for Countering Hate Speech. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 4523–4536.

[71] Eftekhar Hossain, Omar Sharif, Mohammed Moshiul Hoque, and Sarah Masud Preum. 2024. Deciphering Hate: Identifying Hateful Memes and Their Targets. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8347–8359.

[72] Xiaoxuan Hou, Jiayi Yuan, Joel Z. Leibo, and Natasha Jaques. 2024. InvestESG: A Multi-Agent Reinforcement Learning Benchmark for Studying Climate Investment as a Social Dilemma. In *The Thirteenth International Conference on Learning Representations*.

[73] Tiancheng Hu and Nigel Collier. 2024. Quantifying the Persona Effect in LLM Simulations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10289–10307.

[74] Yuxue Hu, Junsong Li, Mingmin Wu, Zhongqiang Huang, Gang Chen, and Ying Sha. 2024. Uncovering and Mitigating the Hidden Chasm: A Study on the Text-Text Domain Gap in Euphemism Identification. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 16 (2024), 18270–18278.

[75] Yangliu Hu, Zikai Song, Na Feng, Yawei Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2025. Sf2t: Self-supervised fragment finetuning of video-llms for fine-grained understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 29108–29117.

[76] Zhe Hu, Hou Pong Chan, Jing Li, and Yu Yin. 2025. Debate-to-Write: A Persona-Driven Multi-Agent Framework for Diverse Argument Generation. In *Proceedings of the 31st International Conference on Computational Linguistics*. 4689–4703.

[77] Yuncheng Hua, Lizhen Qu, and Reza Haf. 2024. Assistive Large Language Model Agents for Socially-Aware Negotiation Dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 8047–8074.

[78] Feng Huang, Qiang Huang, YueTong Zhao, ZhiXiao Qi, BingKun Wang, YongFeng Huang, and SongBin Li. 2024. A Three-Stage Framework for Event-Event Relation Extraction with Large Language Model. In *Neural Information Processing*, Vol. 1968. 434–446.

[79] Kung-Hsiang Huang, Kathleen McKeown, Preslav Nakov, Yejin Choi, and Heng Ji. 2023. Faking Fake News for Real Fake News Detection: Propaganda-Loaded Training Data Generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14571–14589.

[80] Zhongquan Jian, Ante Wang, Jinsong Su, Junfeng Yao, Meihong Wang, and Qingqiang Wu. 2024. EmoTrans: Emotional Transition-Based Model for Emotion Recognition in Conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 5723–5733.

[81] Pengfei Jiao, Hongqian Chen, Qing Bao, Wang Zhang, and Huaming Wu. 2024. Enhancing Multi-Scale Diffusion Prediction via Sequential Hypergraphs and Adversarial Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 8 (2024), 8571–8581.

[82] Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. Persuading across Diverse Domains: A Dataset and Persuasion Large Language Model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1678–1706.

[83] Yiqiao Jin, Minje Choi, Gaurav Verma, Jindong Wang, and Srijan Kumar. 2024. MM-SOC: Benchmarking Multimodal Large Language Models in Social Media Platforms. In *Findings of the Association for Computational Linguistics ACL 2024*. 6192–6210.

[84] David Jurgens, Agrima Seth, Jackson Sargent, Athena Aghighi, and Michael Geraci. 2023. Your Spouse Needs Professional Help: Determining the Contextual Appropriateness of Messages through Modeling Social Relationships. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10994–11013.

[85] Danial Kamali, Joseph D. Romain, Huiyi Liu, Wei Peng, Jingbo Meng, and Parisa Kordjamshidi. 2024. Using Persuasive Writing Strategies to Explain and Detect Health Misinformation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 17285–17309.

[86] Dongjun Kang, Joonsuk Park, Yohan Jo, and JinYeong Bak. 2023. From Values to Opinions: Predicting Human Behaviors and Stances Using Value-Injected Large Language Models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[87] Kateryna Kasianenko, Shima Khanehzar, Stephen Wan, Ehsan Dehghan, and Axel Bruns. 2024. Detecting Online Community Practices with Large Language Models: A Case Study of Pro-Ukrainian Publics on Twitter. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 20106–20135.

[88] Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. 2025. Multiple LLM Agents Debate for Equitable Cultural Alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 24841–24877.

[89] Jaehong Kim, Chaeyoon Jeong, Seongchan Park, Meeyoung Cha, and Wonjae Lee. 2024. How Do Moral Emotions Shape Political Participation? A Cross-Cultural Analysis of Online Petitions Using Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*. 16274–16289.

[90] Sein Kim, Namkyeong Lee, Donghyun Kim, Minchul Yang, and Chanyoung Park. 2023. Task Relation-aware Continual User Representation Learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '23)*. 1107–1119.

[91] Dayoon Ko, Sangho Lee, and Gunhee Kim. 2023. Can Language Models Laugh at YouTube Short-Form Videos?. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[92] Aobo Kong, Wentao Ma, Shiwan Zhao, Yongbin Li, Yuchuan Wu, Ke Wang, Xiaoqian Liu, Qicheng Li, Yong Qin, et al. 2025. SDPO: Segment-Level Direct Preference Optimization for Social Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12409–12423.

[93] Minseo Koo, Doeun Kim, Sungwon Han, and Sungkyu Shaun Park. 2024. Platform-Invariant Topic Modeling via Contrastive Learning to Mitigate Platform-Induced Bias. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 11123–11139.

[94] Julia Kruk, Michela Marchini, Rijul Magu, Caleb Ziems, David Muchlinski, and Diyi Yang. 2024. Silent Signals, Loud Impact: LLMs for Word-Sense Disambiguation of Coded Dog Whistles. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12493–12509.

[95] Marino Kühne, Panagiotis D. Grontas, Giulia De Pasquale, Giuseppe Belgioioso, Florian Dorfler, and John Lygeros. 2025. Optimizing Social Network Interventions via Hypergradient-Based Recommender System Design. In *Forty-Second International Conference on Machine Learning*.

[96] Raghvendra Kumar, Mohammed Salman S A, Aryan Sahu, Tridib Nandi, Pragathi Y P, Sriparna Saha, and Jose G Moreno. 2025. COSMMIC: Comment-Sensitive Multimodal Multilingual Indian Corpus for Summarization and Headline Generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 8728–8748.

[97] Gitanjali Kumari, Kirtan Jain, and Asif Ekbal. 2024. M3Hop-CoT: Misogynous Meme Identification with Multimodal Multi-hop Chain-of-Thought. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 22105–22138.

[98] Bolin Lai, Hongxin Zhang, Miao Liu, Aryan Pariani, Fiona Ryan, Wenqi Jia, Shirley Anugrah Hayati, James Rehg, and Diyi Yang. 2023. Werewolf Among Us: Multimodal Resources for Modeling Persuasion Behaviors in Social Deduction Games. In *Findings of the Association for Computational Linguistics: ACL 2023*. 6570–6588.

[99] Xiaochong Lan, Chen Gao, Depeng Jin, and Yong Li. 2024. Stance Detection with Collaborative Role-Infused LLM-Based Agents. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (2024), 891–903.

[100] Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 128–145.

[101] Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. 2025. Retrieval-Augmented Dynamic Prompt Tuning for Incomplete Multimodal Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 17 (2025), 18035–18043.

[102] Junghun Lee, Hyunju Kim, Fanchen Bu, Jihoon Ko, and Kijung Shin. 2025. DiffIM: Differentiable Influence Minimization with Surrogate Modeling and Continuous Relaxation. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 17 (2025), 18128–18136.

[103] Yuanyuan Lei and Ruihong Huang. 2023. Identifying Conspiracy Theories News Based on Event Relation Graph. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[104] Chin-wing Leung, Tom Lenaerts, and Paolo Turrini. 2024. To Promote Full Cooperation in Social Dilemmas, Agents Need to Unlearn Loyalty. In *Proceedings of the Thirty-ThirdInternational Joint Conference on Artificial Intelligence*, Vol. 1. 111–119.

[105] Ang Li, Bin Liang, Jingqian Zhao, Bowen Zhang, Min Yang, and Ruifeng Xu. 2023. Stance Detection on Social Media with Background Knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 15703–15717.

[106] Hao Li, Ruoyuan Gong, and Hao Jiang. 2025. Political Actor Agent: Simulating Legislative System for Roll Call Votes Prediction with Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 1 (2025), 388–396.

[107] Junliang Li, Yajun Yang, Yujia Zhang, Qinghua Hu, Alan Zhao, and Hong Gao. 2025. Public Opinion Field Effect and Hawkes Process Join Hands for Information Popularity Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 11 (2025), 12076–12083.

[108] Keyu Li, Jin Gao, and Dequan Wang. 2025. Aligned Agents, Biased Swarm: Measuring Bias Amplification in Multi-Agent Systems. In *The Fourteenth International Conference on Learning Representations*.

[109] Miao Li, Jey Han Lau, Eduard Hovy, and Mirella Lapata. 2025. Decomposed Opinion Summarization with Verified Aspect-Aware Modules. In *Findings of the Association for Computational Linguistics: ACL 2025*. 24805–24841.

[110] Wei Li, Jiawen Deng, Jiali You, Yuanyuan He, Yan Zhuang, and Fuji Ren. 2025. ETS-MM: A Multi-Modal Social Bot Detection Model Based on Enhanced Textual Semantic Representation. In *THE WEB CONFERENCE 2025*.

[111] Wenbing Li, Zikai Song, Jielei Zhang, Tianhao Zhao, Junkai Lin, Yiran Wang, and Wei Yang. 2026. Large Language Model as Token Compressor and Decompressor. arXiv:2603.25340 [cs.CL]

[112] Wenbing Li, Zikai Song, Hang Zhou, Yunyao Zhang, Junqing Yu, and Wei Yang. 2025. LoRA-Mixer: Coordinate Modular LoRA Experts Through Serial Attention Routing. *arXiv preprint arXiv:2507.00029* (2025).

[113] Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. 2024. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Advances in Neural Information Processing Systems* 37 (2024), 59808–59832.

[114] Xiang Li, Yucheng Zhou, Laiping Zhao, Jing Li, and Fangming Liu. 2025. Impromptu Cybercrime Euphemism Detection. In *Proceedings of the 31st International Conference on Computational Linguistics*. 9112–9123.

[115] Yingjie Li, Krishna Garg, and Cornelia Caragea. 2023. A New Direction in Stance Detection: Target-Stance Extraction in the Wild. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10071–10085.

[116] Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Zhilin Wang, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2024. MAGE: Machine-generated Text Detection in the Wild. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 36–53.

[117] Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-Modal Stance Detection: New Datasets and Model. In *Findings of the Association for Computational Linguistics ACL 2024*. 12373–12387.

[118] Jian Liao, Yu Feng, Yujin Zheng, Jun Zhao, Suge Wang, and JianXing Zheng. 2025. My Words Imply Your Opinion: Reader Agent-Based Propagation Enhancement for Personalized Implicit Emotion Analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 16156–16172.

[119] Hongzhan Lin, Pengyao Yi, Jing Ma, Haiyun Jiang, Ziyang Luo, Shuming Shi, and Ruifang Liu. 2023. Zero-Shot Rumor Detection with Propagation Structure via Prompt Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 4 (2023), 5213–5221.

[120] Bo Ling, Yan Lyu, Dongxiao Li, Guanyu Gao, Yi Shi, Xueyong Xu, and Weiwei Wu. 2024. SocialGAIL: Faithful Crowd Simulation for Social Robot Navigation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 16873–16880.

[121] Jintao Liu, Zequn Zhang, Kaiwen Wei, Zhi Guo, Xian Sun, Li Jin, and Xiaoyu Li. 2023. Event Causality Extraction via Implicit Cause-Effect Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 6792–6804.

[122] Panfeng Liu, Guoliang Qiu, Biaoshuai Tao, and Kuan Yang. 2025. A Thorough Comparison Between Independent Cascade and Susceptible-Infected-Recovered Models. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 1 (2025), 487–495.

[123] Xiawei Liu, Shiyue Yang, Xinnong Zhang, Haoyu Kuang, Libo Sun, Yihang Yang, Siming Chen, Xuanjing Huang, and Zhongyu Wei. 2025. AI-press: A Multi-Agent News Generating and Feedback Simulation System Powered by Large Language Models. In *Proceedings of the 31st International Conference on Computational Linguistics: System Demonstrations*. 63–82.

[124] Yanchen Liu, Mingyu Derek Ma, Wenna Qin, Azure Zhou, Jiaao Chen, Weiyan Shi, Wei Wang, and Diyi Yang. 2024. Decoding Susceptibility: Modeling Misbelief to Misinformation Through a Computational Approach. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 15178–15194.

[125] Yuhan Liu, Zirui Song, Juntian Zhang, Xiaoqing Zhang, Xiuying Chen, and Rui Yan. 2025. The Stepwise Deception: Simulating the Evolution from True News to Fake News with LLM Agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 26187–26203.

[126] Yanchao Liu, Pengzhou Zhang, Wenchao Song, Yao Zheng, Deyu Li, Lei Shi, and Junpeng Gong. 2025. THGNets: Constrained Temporal Hypergraphs and Graph Neural Networks in Hyperbolic Space for Information Diffusion Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 12 (2025), 12220–12228.

[127] Ziyan Liu, Chunxiao Fan, Haoran Lou, Yuexin Wu, and Kaiwei Deng. 2025. MIND: A Multi-agent Framework for Zero-shot Harmful Meme Detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 923–947.

[128] Meng Luo, Hao Fei, Bobo Li, Shengqiong Wu, Qian Liu, Soujanya Poria, Erik Cambria, Mong-Li Lee, and Wynne Hsu. 2024. PanoSent: A Panoptic Sextuple Extraction Benchmark for Multimodal Conversational Aspect-based Sentiment Analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 7667–7676.

[129] Mingjin Lv, Zelin Liu, Feifei Shao, Yi-Ping Phoebe Chen, Junqing Yu, Wei Yang, and Zikai Song. 2026. GateMOT: Q-Gated Attention for Dense Object Tracking. *arXiv preprint arXiv:2604.26353* (2026).

[130] Jing Ma, Chen Chen, Chunyan Hou, and Xiaojie Yuan. 2023. KAPALM: Knowledge grAPh enhAnced Language Models for Fake News Detection. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[131] Abdurahman Maarouf, Dominik Bär, Dominique Geissler, and Stefan Feuerriegel. 2024. HQP: A Human-Annotated Dataset for Detecting Online Propaganda. In *Findings of the Association for Computational Linguistics ACL 2024*. 6064–6089.

[132] Dominik Macko, Jakub Kopál, Robert Moro, and Ivan Srba. 2025. MultiSocial: Multilingual Benchmark of Machine-Generated Text Detection of Social-Media Texts. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 727–752.

[133] Khyati Mahajan and Samira Shaikh. 2024. Persona-Aware Multi-Party Conversation Response Generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 12712–12723.

[134] Fanhang Man, Huandong Wang, Jianjie Fang, Zhaoyi Deng, Baining Zhao, Xinlei Chen, and Yong Li. 2025. Context-Aware Sentiment Forecasting via LLM-based Multi-Perspective Role-Playing Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2687–2703.

[135] Mounika Marreddy, Subba Reddy Oota, Venkata Charan Chinni, Manish Gupta, and Lucie Flek. 2025. USDC: A Dataset of user stance and dogmatism in Long conversations. In *Findings of the Association for Computational Linguistics: ACL 2025*. 23715–23759.

[136] Stan Matwin, Aristides Milios, Paweł Prałat, Amilcar Soares, and François Théberge. 2021. *Survey of Generative Methods for Social Media Analysis*. arXiv:2112.07041

[137] Jingbiao Mei, Jinghong Chen, Weizhe Lin, Bill Byrne, and Marcus Tomalin. 2024. Improving Hateful Meme Detection through Retrieval-Guided Contrastive Learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5333–5347.

[138] Julia Mendelsohn, Ronan Le Bras, Yejin Choi, and Maarten Sap. 2023. From Dogwhistles to Bullhorns: Unveiling Coded Rhetoric with Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15162–15180.

[139] Changrong Min, Ximing Li, Liang Yang, Zhilin Wang, Bo Xu, and Hongfei Lin. 2023. Just Like a Human Would, Direct Access to Sarcasm Augmented with Potential Result and Reaction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 10172–10183.

[140] Kentaro Miyake, Hiroyoshi Ito, Christos Faloutsos, Hirotomo Matsumoto, and Atsuyuki Morishima. 2024. NETEVOLVE: Social Network Forecasting Using Multi-Agent Reinforcement Learning with Interpretable Features. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*. 2542–2551.

[141] Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, et al. 2026. From Individual to Society: A Survey on Social Simulation Driven by Large Language Model-Based Agents. *ACM Comput. Surv.* 58, 11 (2026), 280:1–280:41.

[142] Xinyi Mou, Zhongyu Wei, and Xuanjing Huang. 2024. Unveiling the Truth and Facilitating Change: Towards Agent-based Large-scale Social Movement Simulation. In *Findings of the Association for Computational Linguistics ACL 2024*. 4789–4809.

[143] Yida Mu, Mali Jin, Kalina Bontcheva, and Xingyi Song. 2024. Examining Temporalities on Stance Detection towards COVID-19 Vaccination. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 6732–6738.

[144] Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Said Ahmad, Meriem Beloucif, Saif M. Mohammad, et al. 2023. AfriSenti: A Twitter Sentiment Analysis Benchmark for African Languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[145] Sankha Subhra Mullick, Mohan Bhambhani, Suhit Sinha, Akshat Mathur, Somya Gupta, and Jidnya Shah. 2023. Content Moderation for Evolving Policies Using Binary Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*. 561–573.

[146] Sharath Naganna, Saprativa Bhattacharjee, Biplab Banerjee, and Pushpak Bhattacharyya. 2025. "My Life Is Miserable, Have to Sign 500 Autographs Everyday": Exposing Humblebragging, the Brags in Disguise. In *Findings of the Association for Computational Linguistics: ACL 2025*. 3838–3858.

[147] Qiong Nan, Qiang Sheng, Juan Cao, Beizhe Hu, Danding Wang, and Jintao Li. 2024. Let Silence Speak: Enhancing Fake News Detection with Generated Comments from Large Language Models. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. 1732–1742.

[148] Buddhika Nettasinghe, Ashwin Rao, Bohan Jiang, Allon Percus, and Kristina Lerman. 2025. In-Group Love, out-Group Hate: A Framework to Measure Affective Polarization via Contentious Online Discussions. In *THE WEB CONFERENCE 2025*.

[149] Huy Nghiem and Hal Daumé Iii. 2024. HateCOT: An Explanation-Enhanced Dataset for Generalizable Offensive Speech Detection via Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 5938–5956.

[150] Mikel K. Ngueajio, Saurav Aryal, Marcellin Atemkeng, Gloria Washington, and Danda Rawat. 2025. Decoding Fake News and Hate Speech: A Survey of Explainable AI Techniques. *Comput. Surveys* 57, 7 (2025), 1–37.

[151] Dung Nguyen, Phuoc Nguyen, Hung Le, Kien Do, Svetha Venkatesh, and Truyen Tran. 2023. Memory-Augmented Theory of Mind Network. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 10 (2023), 11630–11637.

[152] Khoi P. N. Nguyen and Vincent Ng. 2024. Computational Meme Understanding: A Survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 21251–21267.

[153] Cheng Niu, Yang Guan, Yuanhao Wu, Juno Zhu, Juntong Song, Randy Zhong, Kaihua Zhu, Siliang Xu, Shizhe Diao, et al. 2024. VeraCT Scan: Retrieval-Augmented Fake News Detection with Justifiable Reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. 266–277.

[154] Fuqiang Niu, Min Yang, Ang Li, Baoquan Zhang, Xiaojiang Peng, and Bowen Zhang. 2024. A Challenge Dataset and Effective Models for Conversational Stance Detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 122–132.

[155] Célia Nouri, Jean-Philippe Cointet, and Chloé Clavel. 2025. Graphically Speaking: Unmasking Abuse in Social Media with Conversation Insights. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 18271–18286.

[156] Rajvardhan Oak, Muhammad Haroon, Claire Wonjeong Jo, Magdalena Wojcieszak, and Anshuman Chhabra. 2025. Re-Ranking Using Large Language Models for Mitigating Exposure to Harmful Content on Social Media Platforms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 894–908.

[157] Xianghe Pang, Shuo Tang, Rui Ye, Yuxin Xiong, Bolun Zhang, Yanfeng Wang, and Siheng Chen. 2024. Self-Alignment of Large Language Models via Monopolylogue-Based Social Scene Simulation. In *Proceedings of the 41st International Conference on Machine Learning*. 39416–39447.

[158] Tanmay Parekh, Jeffrey Kwan, Jiarui Yu, Sparsh Johri, Hyosang Ahn, Sreya Muppalla, Kai-Wei Chang, Wei Wang, and Nanyun Peng. 2024. SPEED++: A Multilingual Event Extraction Framework for Epidemic Prediction and Preparedness. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 12936–12965.

[159] Chaewon Park, Soohwan Kim, Kyubyong Park, and Kunwoo Park. 2023. K-HATERS: A Hate Speech Detection Corpus in Korean with Target-Specific Ratings. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[160] Chan Young Park, Shuyue Stella Li, Hayoung Jung, Svitlana Volkova, Tanu Mitra, David Jurgens, and Yulia Tsvetkov. 2024. ValueScope: Unveiling Implicit Norms and Values via Return Potential Model of Social Interactions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 16659–16695.

[161] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. *Generative Agents: Interactive Simulacra of Human Behavior.* arXiv:2304.03442

[162] Someen Park, Jaehoon Kim, Seungwan Jin, Sohyun Park, and Kyungsik Han. 2024. PREDICT: Multi-Agent-based Debate Simulation for Generalized Hate Speech Detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 20963–20987.

[163] Hao Peng, Ruitong Zhang, Shaoning Li, Yuwei Cao, Shirui Pan, and Philip S. Yu. 2023. Reinforced, Incremental and Cross-Lingual Event Detection From Social Messages. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 1 (2023), 980–998.

[164] Viet Thanh Pham, Lizhen Qu, Zhuang Li, Suraj Sharma, and Gholamreza Haffari. 2025. SurveyPilot: An Agentic Framework for Automated Human Opinion Collection from Social Media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4397–4422.

[165] Khiem Phi, Noushin Salek Faramarzi, Chenlu Wang, and Ritwik Banerjee. 2024. Paying Attention to Deflections: Mining Pragmatic Nuances for Whataboutism Detection in Online Discourse. In *Findings of the Association for Computational Linguistics ACL 2024*. 12628–12643.

[166] Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, et al. 2025. *AgentSociety: Large-scale Simulation of LLM-driven Generative Agents Advances Understanding of Human Behaviors and Society.* doi:10.2139/ssrn.5954414

[167] Flor Miriam Plaza Del Arco, Amanda Curry, Alba Cercas Curry, Gavin Abercrombie, and Dirk Hovy. 2024. Angry Men, Sad Women: Large Language Models Reflect Gendered Stereotypes in Emotion Attribution. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 7682–7696.

[168] Mohammad Mahdi Abdollah Pour, Parsa Farinneya, Manasa Bharadwaj, Nikhil Verma, Ali Pesaranghader, and Scott Sanner. 2023. COUNT: COntrastive UNlikelihood Text Style Transfer for Text Detoxification. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[169] Peng Qi, Yuyan Bu, Juan Cao, Wei Ji, Ruihao Shui, Junbin Xiao, Danding Wang, and Tat-Seng Chua. 2023. FakeSV: A Multimodal Benchmark with Rich Social Context for Fake News Detection on Short Video Platforms. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 12 (2023), 14444–14452.

[170] Boyu Qiao, Kun Li, Wei Zhou, Shilong Li, Qianqian Lu, and Songlin Hu. 2025. BotSim: LLM-Powered Malicious Social Botnet Simulation. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 13 (2025), 14377–14385.

[171] Boyu Qiao, Wei Zhou, Kun Li, Shilong Li, and Songlin Hu. 2025. Dispelling the Fake: Social Bot Detection Based on Edge Confidence Evaluation. *IEEE Transactions on Neural Networks and Learning Systems* 36, 4 (2025), 7302–7315.

[172] Tingrui Qiao, Caroline Walker, Chris W. Cunningham, and Yun Sing Koh. 2025. Thematic-LM: A LLM-based Multi-Agent System for Large-Scale Thematic Analysis. In *THE WEB CONFERENCE 2025*.

[173] Yang Qiao, Liqiang Jing, Xuemeng Song, Xiaolin Chen, Lei Zhu, and Liqiang Nie. 2023. Mutual-Enhanced Incongruity Learning Network for Multi-Modal Sarcasm Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 8 (2023), 9507–9515.

[174] Ziqi Qiu, Jianxing Yu, Yufeng Zhang, Hanjiang Lai, Yanghui Rao, Qinliang Su, and Jian Yin. 2025. Detecting Emotional Incongruity of Sarcasm by Commonsense Reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*. 9062–9073.

[175] Chao Qu, Xiaoyu Tan, Siqiao Xue, Xiaoming Shi, James Zhang, and Hongyuan Mei. 2023. Bellman Meets Hawkes: Model-Based Reinforcement Learning via Temporal Point Processes. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 8 (2023), 9543–9551.

[176] Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. BOOKWORLD: From Novels to Interactive Agent Societies for Story Creation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 15898–15912.

[177] Qibing Ren, Zhijie Zheng, Jiaxuan Guo, Junchi Yan, Lizhuang Ma, and Jing Shao. 2025. When AI Agents Collude Online: Financial Fraud Risks by Collaborative LLM Agents on Social Platforms. In *The Fourteenth International Conference on Learning Representations*.

[178] Daniel Russo, Shane Peter Kaszefski-Yaschuk, Jacopo Staiano, and Marco Guerini. 2023. Countering Misinformation via Emotional Response Generation. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[179] Philipp Seeberger, Dominik Wagner, and Korbinian Riedhammer. 2024. MMUTF: Multimodal Multimedia Event Argument Extraction with Unified Template Filling. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 6539–6548.

[180] Ruiting Shao, Ryan Schwarz, Christopher Clifton, and Edward Delp. 2024. A Natural Approach for Synthetic Short-Form Text Analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 1034–1042.

[181] Shivam Sharma, Siddhant Agarwal, Tharun Suresh, Preslav Nakov, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. What Do You MEME? Generating Explanations for Visual Semantic Role Labelling in Memes. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 8 (2023), 9763–9771.

[182] Shivam Sharma, Ramaneswaran S, Udit Arora, Md. Shad Akhtar, and Tanmoy Chakraborty. 2023. MEMEX: Detecting Explanatory Evidence for Memes via Knowledge-Enriched Contextualization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5272–5290.

[183] Chao Shen, Zhenyu Song, Pengyu He, Limin Liu, and Zhenyu Xiong. 2024. Online Rumors during the COVID-19 Pandemic: Co-Evolution of Themes and Emotions. *Frontiers in Public Health* 12 (2024), 1375731.

[184] Jocelyn Shen, Yubin Kim, Mohit Hulse, Wazeer Zulfikar, Sharifa Alghowinem, Cynthia Breazeal, and Hae Park. 2024. EmpathicStories++: A Multimodal Dataset for Empathy Towards Personal Experiences. In *Findings of the Association for Computational Linguistics ACL 2024*. 4525–4536.

[185] Haojun Shi, Suyu Ye, Xinyu Fang, Chuanyang Jin, Leyla Isik, Yen-Ling Kuo, and Tianmin Shu. 2025. MuMA-ToM: Multi-modal Multi-Agent Theory of Mind. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 2 (2025), 1510–1519.

[186] Yuanchen Shi and Fang Kong. 2024. Integrating Stickers into Multimodal Dialogue Summarization: A Novel Dataset and Approach for Enhancing Social Media Interaction. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 9525–9534.

[187] Jonathan Skaggs, Michael Richards, Melissa Morris, Michael A. Goodrich, and Jacob W. Crandall. 2024. Fostering Collective Action in Complex Societies Using Community-Based Agents. In *Thirty-Third International Joint Conference on Artificial Intelligence*, Vol. 1. 211–219.

[188] Shano Solanki, Mukesh Kumar, and Rakesh Kumar. 2025. A Survey on Information Diffusion and Competitive Influence Maximization in Social Networks. *Social Network Analysis and Mining* 15, 1 (2025), 41.

[189] Jiayu Song, Mahmud Elahi Akhter, Dana Atzil-Slonim, and Maria Liakata. 2025. Temporal Reasoning for Timeline Summarisation in Social Media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 28085–28101.

[190] Maojia Song, Tej Deep Pala, Ruiwen Zhou, Weisheng Jin, Amir Zadeh, Chuan Li, Dorien Herremans, and Soujanya Poria. 2025. Measuring and Mitigating Rapport Bias of Large Language Models under Multi-Agent Social Interactions. In *The Fourteenth International Conference on Learning Representations*.

[191] Xiaoying Song, Sujana Mamidisetty, Eduardo Blanco, and Lingzi Hong. 2025. Assessing the Human Likeness of AI-generated Counterspeech. In *Proceedings of the 31st International Conference on Computational Linguistics*. 3547–3559.

[192] Zikai Song, Run Luo, Lintao Ma, Ying Tang, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. 2025. Temporal Coherent Object Flow for Multi-Object Tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 6978–6986.

[193] Zikai Song, Run Luo, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2023. Compact transformer tracker with correlative masked modeling. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 37. 2321–2329.

[194] Zikai Song, Ying Tang, Run Luo, Lintao Ma, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2024. Autogenic language embedding for coherent point tracking. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 2021–2030.

[195] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. 2022. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8791–8800.

[196] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang, and Xinchao Wang. 2026. Hypergraph-State Collaborative Reasoning for Multi-Object Tracking. arXiv:2604.12665 [cs.CV]

[197] Taylor Sorensen, Liwei Jiang, Jena D. Hwang, Sydney Levine, Valentina Pyatkin, Peter West, Nouha Dziri, Ximing Lu, Kavel Rao, et al. 2024. Value Kaleidoscope: Engaging AI with Pluralistic Human Values, Rights, and Duties. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 18 (2024), 19937–19947.

[198] Dipankar Srirag, Aditya Joshi, Jordan Painter, and Diptesh Kanojia. 2025. BESSTIE: A Benchmark for Sentiment and Sarcasm Classification for Varieties of English. In *Findings of the Association for Computational Linguistics: ACL 2025*. 8413–8429.

[199] Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[200] Haoxin Sun and Zhongzhi Zhang. 2023. Opinion Optimization in Directed Social Networks. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 4 (2023), 4623–4632.

[201] Yuchong Sun, Bei Liu, Xu Chen, Ruihua Song, and Jianlong Fu. 2024. ViCo: Engaging Video Comment Generation with Human Preference Rewards. In *Proceedings of the 6th ACM International Conference on Multimedia in Asia (MMAsia '24)*. 1.

[202] Zhen Sun, Zongmin Zhang, Xinyue Shen, Ziyi Zhang, Yule Liu, Michael Backes, Yang Zhang, and Xinlei He. 2025. Are We in the AI-Generated Text World Already? Quantifying and Monitoring AIGT on Social Media. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 22975–23005.

[203] Shahbaz Syed, Dominik Schwabe, Khalid Al Khatib, and Martin Potthast. 2023. Indicative Summarization of Long Discussions. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[204] Zhaoxuan Tan, Shangbin Feng, Melanie Sclar, Herun Wan, Minnan Luo, Yejin Choi, and Yulia Tsvetkov. 2023. BotPercent: Estimating Bot Populations in Twitter Communities. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[205] Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. Enhancing Personalized Dialogue Generation with Contrastive Latent Variables: Combining Sparse and Dense Persona. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 5456–5468.

[206] Maksym Taranukhin, Vered Shwartz, and Evangelos Milios. 2024. Stance Reasoner: Zero-shot Stance Detection on Social Media with Explicit Reasoning. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 15257–15272.

[207] Ravi Tejwani, Yen-Ling Kuo, Tianmin Shu, Boris Katz, and Andrei Barbu. 2022. Social Interactions as Recursive MDPs. In *Proceedings of the 5th Conference on Robot Learning*. 949–958.

[208] Ravi Tejwani, Yen-Ling Kuo, Tianmin Shu, Bennett Stankovits, Dan Gutfreund, Joshua B. Tenenbaum, Boris Katz, and Andrei Barbu. 2022. Incorporating Rich Social Interactions Into MDPs. In *2022 International Conference on Robotics and Automation (ICRA)*. 7395–7401.

[209] Ravi Tejwani, Yen-Ling Kuo, Tianmin Shu, Bennett Stankovits, Dan Gutfreund, Joshua B. Tenenbaum, Boris Katz, and Andrei Barbu. 2023. Zero-Shot Linear Combinations of Grounded Social Interactions with Linear Social MDPs. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 1 (2023), 94–101.

[210] Elizaveta Tennant, Stephen Hailes, and Mirco Musolesi. 2024. Moral Alignment for LLM Agents. In *The Thirteenth International Conference on Learning Representations*.

[211] Surendrabikram Thapa, Shuvam Shiwakoti, Siddhant Bikram Shah, Surabhi Adhikari, Hariram Veeramani, Mehwish Nasim, and Usman Naseem. 2025. Large Language Models (LLM) in Computational Social Science: Prospects, Current State, and Challenges. *Social Network Analysis and Mining* 15, 1 (2025), 4.

[212] Chong Tian, Qirong Ho, and Xiuying Chen. 2025. A Symbolic Adversarial Learning Framework for Evolving Fake News Generation and Detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. 12307–12321.

[213] Balaji T.K., Chandra Sekhara Rao Annavarapu, and Annushree Bablani. 2021. Machine Learning Algorithms for Social Media Analysis: A Survey. *Computer Science Review* 40 (2021), 100395.

[214] Manuel Tonneau, Diyi Liu, Niyati Malhotra, Scott A. Hale, Samuel Fraiberger, Victor Orozco-Olvera, and Paul Röttger. 2025. HateDay: Insights from a Global Hate Speech Dataset Representative of a Day on Twitter. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2297–2321.

[215] Pranav Narayanan Venkit, Mukund Srinath, Sanjana Gautam, Saranya Venkatraman, Vipul Gupta, Rebecca J. Passonneau, and Shomir Wilson. 2023. The Sentiment Problem: A Critical Survey towards Deconstructing Sentiment Analysis. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[216] Herun Wan, Minnan Luo, Zhixiong Su, Guang Dai, and Xiang Zhao. 2025. On the Risk of Evidence Pollution for Malicious Social Text Detection in the Era of LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9731–9761.

[217] Bing Wang, Shengsheng Wang, Changchun Li, Renchu Guan, and Ximing Li. 2024. Harmfully Manipulated Images Matter in Multimodal Misinformation Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 2262–2271.

[218] Chenxu Wang, Bin Dai, Huaping Liu, and Baoyuan Wang. 2024. Towards Objectively Benchmarking Social Intelligence of Language Agents at the Action Level. In *Findings of the Association for Computational Linguistics ACL 2024*. 8885–8897.

[219] Chenxi Wang, Zongfang Liu, Dequan Yang, and Xiuying Chen. 2025. Decoding Echo Chambers: LLM-powered Simulations Revealing Polarization in Social Networks. In *Proceedings of the 31st International Conference on Computational Linguistics*. 3913–3923.

[220] Dali Wang, Yunyao Zhang, Junqing Yu, Yi-Ping Phoebe Chen, Chen Xu, and Zikai Song. 2026. Seeing Further and Wider: Joint Spatio-Temporal Enlargement for Micro-Video Popularity Prediction. *arXiv preprint arXiv:2604.20311* (2026).

[221] Haiyang Wang, Yuchen Pan, Xin Song, Xuechen Zhao, Minghao Hu, and Bin Zhou. 2024. F2RL: Factuality and Faithfulness Reinforcement Learning Framework for Claim-Guided Evidence-Supported Counterspeech Generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 4457–4470.

[222] Han Wang, Tan Rui Yang, Usman Naseem, and Roy Ka-Wei Lee. 2024. MultiHateClip: A Multilingual Benchmark Dataset for Hateful Video Detection on YouTube and Bilibili. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 7493–7502.

[223] Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024. SOTOPIA-\ensuremath\pi: Interactive Learning of Socially Intelligent Language Agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers). 12912–12940.

[224] Yiding Wang, Yuxuan Chen, Fangwei Zhong, Long Ma, and Yizhou Wang. 2024. Simulating Human-like Daily Activities with Desire-Driven Autonomy. In *The Thirteenth International Conference on Learning Representations*.

[225] Yan Wang, Bo Wang, Yachao Zhao, Dongming Zhao, Xiaojia Jin, Jijun Zhang, Ruifang He, and Yuexian Hou. 2024. Emotion Recognition in Conversation via Dynamic Personality. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 5711–5722.

[226] Yuxin Wang, Ivory Yang, Saeed Hassanpour, and Soroush Vosoughi. 2024. MentalManip: A Dataset For Fine-grained Analysis of Mental Manipulation in Conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 3747–3764.

[227] Yujie Wang, Yunwei Zhao, Jing Yang, Han Han, Shiguang Shan, and Jie Zhang. 2025. *Evaluating Cognitive-Behavioral Fixation via Multimodal User Viewing Patterns on Social Media*. arXiv:2509.04823

[228] Zhen Wang, Zhao Song, Chen Shen, and Shuyue Hu. 2023. Emergence of Punishment in Social Dilemma with Environmental Feedback. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 10 (2023), 11708–11716.

[229] Yiluo Wei and Gareth Tyson. 2024. Understanding the Impact of AI-Generated Content on Social Media: The Pixiv Case. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 6813–6822.

[230] Haoyang Wen, Zhenxin Xiao, Eduard Hovy, and Alexander Hauptmann. 2023. Towards Open-Domain Twitter User Profile Inference. In *Findings of the Association for Computational Linguistics: ACL 2023*. 3172–3188.

[231] Chenhao Wu, Ruifang He, Chang Liu, and Bo Wang. 2024. Continuous Relational Diffusion Driven Topic Model with Multi-Grained Text for Microblog. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 3897–3906.

[232] Daiqing Wu, Dongbao Yang, Yu Zhou, and Can Ma. 2024. Robust Multimodal Sentiment Analysis of Image-Text Pairs by Distribution-Based Feature Recovery and Fusion. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 5780–5789.

[233] Kun Wu, Xinyi Mou, Lanqing Xue, Zhenzhe Ying, Weiqiang Wang, Qi Zhang, Xuanjing Huang, and Zhongyu Wei. 2024. PASUM: A Pre-Training Architecture for Social Media User Modeling Based on Text Graph. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 12644–12656.

[234] Wayne Wu, Honglin He, Jack He, Yiran Wang, Chenda Duan, Zhizheng Liu, Quanyi Li, and Bolei Zhou. 2024. MetaUrban: An Embodied AI Simulation Platform for Urban Micromobility. In *The Thirteenth International Conference on Learning Representations*.

[235] Wen Wu, Bo Li, Chao Zhang, Chung-Cheng Chiu, Qiujia Li, Junwen Bai, Tara Sainath, and Phil Woodland. 2024. Handling Ambiguity in Emotion: From Out-of-Domain Detection to Distribution Estimation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2078–2093.

[236] Yafeng Wu, Yunyao Zhang, Liliang Ye, Guiyi Zeng, Junqing Yu, Chen Xu, and Zikai Song. 2026. HotComment: A Benchmark for Evaluating Popularity of Online Comments. *arXiv preprint arXiv:2604.25614* (2026).

[237] Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian I. Kwon, Makoto Onizuka, Shaojie Tang, and Chuan Xiao. 2024. Shall We Team Up: Exploring Spontaneous Cooperation of Competing LLM Agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 5163–5186.

[238] Yantuan Xian, Pu Li, Hao Peng, Zhengtao Yu, Yan Xiang, and Philip S. Yu. 2025. Community Detection in Large-Scale Complex Networks via Structural Entropy Game. In *THE WEB CONFERENCE 2025*.

[239] Yang Xiao, Jiashuo Wang, Qiancheng Xu, Changhe Song, Chunpu Xu, Yi Cheng, Wenjie Li, and Pengfei Liu. 2025. Towards Dynamic Theory of Mind: Evaluating LLM Adaptation to Temporal Evolution of Human States. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 24036–24057.

[240] Jiadong Xie, Fan Zhang, Kai Wang, Xuemin Lin, and Wenjie Zhang. 2023. Minimizing the Influence of Misinformation via Vertex Blocking. In *2023 IEEE 39th International Conference on Data Engineering (ICDE)*. 789–801.

[241] Haowei Xu, Chao Gao, Xianghua Li, and Zhen Wang. 2025. HyperIDP: Customizing Temporal Hypergraph Neural Networks for Multi-Scale Information Diffusion Prediction. In *Proceedings of the 31st International Conference on Computational Linguistics*. 964–977.

[242] Kele Xu, Qisheng Xu, Binli Luo, Han Zhou, Zengming Lin, Hui Geng, and Xianhan Tan. 2025. Higher-Order Vision-Language Fusion for Video Popularity Prediction. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*. 14086–14093.

[243] Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, et al. 2025. When Agents "Misremember" Collectively: Exploring the Mandela Effect in LLM-based Multi-Agent Systems. In *The Fourteenth International Conference on Learning Representations*.

[244] Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The Earth Is Flat Because...: Investigating LLMs' Belief towards Misinformation via Persuasive Conversation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 16259–16303.

[245] Ruoxi Xu, Yingfei Sun, Mengjie Ren, Shiguang Guo, Ruotong Pan, Hongyu Lin, Le Sun, and Xianpei Han. 2024. AI for Social Science and Social Science of AI: A Survey. *Information Processing & Management* 61, 3 (2024), 103665.

[246] Yihuai Xu, Yongwei Wang, Yifei Bi, Huangsen Cao, Zhouhan Lin, Yu Zhao, and Fei Wu. 2024. Training-Free LLM-generated Text Detection by Mining Token Probability Sequences. In *The Thirteenth International Conference on Learning Representations*.

[247] Yongxiu Xu, Hao Xu, Heyan Huang, Shiyao Cui, Minghao Tang, Longzheng Wang, and Hongbo Xu. 2024. An Effective Span-Based Multimodal Named Entity Recognition with Consistent Cross-Modal Alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 1063–1072.

[248] Yijie Xu, Bolun Zheng, Wei Zhu, Hangjia Pan, Yuchen Yao, Ning Xu, Anan Liu, Quan Zhang, and Chenggang Yan. 2025. SMTPD: A New Benchmark for Temporal Prediction of Social Media Popularity. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 18847–18857.

[249] Zhenjie Xu, Wenqing Chen, Yi Tang, Xuanying Li, Cheng Hu, Zhixuan Chu, Kui Ren, Zibin Zheng, and Zhichao Lu. 2025. Mitigating Social Bias in Large Language Models: A Multi-Objective Approach Within a Multi-Agent Framework. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 24 (2025), 25579–25587.

[250] Kaiqi Yang, Hang Li, Hongzhi Wen, Tai-Quan Peng, Jiliang Tang, and Hui Liu. 2024. Are Large Language Models (LLMs) Good Social Predictors?. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 2718–2730.

[251] Qu Yang, Qinghongya Shi, Tongxin Wang, and Mang Ye. 2025. Uncertain Multimodal Intention and Emotion Understanding in the Wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 24700–24709.

[252] Ruichao Yang, Wei Gao, Jing Ma, Hongzhan Lin, and Bo Wang. 2024. Reinforcement Tuning for Detecting Stances and Debunking Rumors Jointly with Large Language Models. In *Findings of the Association for Computational Linguistics ACL 2024*. 13423–13439.

[253] Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 2023. Orders Are Unwanted: Dynamic Deep Graph Convolutional Network for Personality Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 11 (2023), 13896–13904.

[254] Xianjun Yang, Liangming Pan, Xuandong Zhao, Haifeng Chen, Linda Ruth Petzold, William Yang Wang, and Wei Cheng. 2024. A Survey on Detection of LLMs-Generated Content. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. 9786–9805.

[255] Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable Hate Speech Detection with Step-by-Step Reasoning. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[256] Yutao Yang, Junsong Li, Qianjun Pan, Bihao Zhan, Yuxuan Cai, Lin Du, Jie Zhou, Kai Chen, Qin Chen, et al. 2026. *AutoSkill: Experience-driven Lifelong Learning via Skill Self-Evolution.* arXiv:2603.01145

[257] Yixin Yang, Zheng Li, Qingxiu Dong, Heming Xia, and Zhifang Sui. 2024. Can Large Multimodal Models Uncover Deep Semantics Behind Images?. In *Findings of the Association for Computational Linguistics ACL 2024*. 1898–1912.

[258] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Jinsong Chen, Martz Ma, et al. 2025. *OASIS: Open Agent Social Interaction Simulations with One Million Agents.* arXiv:2411.11581

[259] Junchi Yao, Hongjie Zhang, Jie Ou, Dingyi Zuo, Zheng Yang, and Zhicheng Dong. 2025. Social Opinions Prediction Utilizes Fusing Dynamics Equation with LLM-based Agents. *Scientific Reports* 15, 1 (2025), 15472.

[260] Yaowenhu, Wenxuan Tu, Yue Liu, Xinhang Wan, Junyi Yan, Taichun Zhou, and Xinwang Liu. 2025. Scalable Attribute-Missing Graph Clustering via Neighborhood Differentiation. In *Forty-Second International Conference on Machine Learning*.

[261] Liliang Ye, Guiyi Zeng, Yunyao Zhang, Yi-Ping Phoebe Chen, Junqing Yu, and Zikai Song. 2026. OmniTrend: Content-Context Modeling for Scalable Social Popularity Prediction. *arXiv preprint arXiv:2604.26252* (2026).

[262] Liliang Ye, Yunyao Zhang, Yafeng Wu, Yi-Ping Phoebe Chen, Junqing Yu, Wei Yang, and Zikai Song. 2025. MVP: Winning Solution to SMP Challenge 2025 Video Track. *arXiv preprint arXiv:2507.00950* (2025).

[263] Qichao Ying, Xiaoxiao Hu, Yangming Zhou, Zhenxing Qian, Dan Zeng, and Shiming Ge. 2023. Bootstrapping Multi-View Representations for Fake News Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 4 (2023), 5384–5392.

[264] Erxin Yu, Jing Li, and Chunpu Xu. 2024. PopALM: Popularity-aligned Language Models for Social Media Trendy Response Prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 12867–12878.

[265] Jianfei Yu, Ziyan Li, Jieming Wang, and Rui Xia. 2023. Grounded Multimodal Named Entity Recognition on Social Media. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 9141–9154.

[266] Xiaoyan Yu, Tongxu Luo, Yifan Wei, Fangyu Lei, Yiming Huang, Hao Peng, and Liehuang Zhu. 2024. Neeko: Leveraging Dynamic LoRA for Efficient Multi-Character Role-Playing Agent. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 12540–12557.

[267] Xiaoyan Yu, Yifan Wei, Shuaishuai Zhou, Zhiwei Yang, Li Sun, Hao Peng, Liehuang Zhu, and Philip S. Yu. 2025. Towards Effective, Efficient and Unsupervised Social Event Detection in the Hyperbolic Space. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 12 (2025), 13106–13114.

[268] Wajdi Zaghouani, Hamdy Mubarak, and Md. Rafiul Biswas. 2024. So Hateful! Building a Multi-Label Hate Speech Annotated Arabic Dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 15044–15055.

[269] Guiyi Zeng, Junqing Yu, Yi-Ping Phoebe Chen, Xu Chen, Wei Yang, and Zikai Song. 2026. CurEvo: Curriculum-Guided Self-Evolution for Video Understanding. *arXiv preprint arXiv:2604.26707* (2026).

[270] Jingjie Zeng, Liang Yang, Zekun Wang, Yuanyuan Sun, and Hongfei Lin. 2025. Sheep's Skin, Wolf's Deeds: Are LLMs Ready for Metaphorical Implicit Hate Speech?. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 16657–16677.

[271] Guibin Zhang, Muxin Fu, and Shuicheng Yan. 2025. MemGen: Weaving Generative Latent Memory for Self-Evolving Agents. In *The Fourteenth International Conference on Learning Representations*.

[272] Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. 2026. *MemSkill: Learning and Evolving Memory Skills for Self-Evolving Agents.* arXiv:2602.02474

[273] Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards Multimodal Empathetic Response Generation: A Rich Text-Speech-Vision Avatar-Based Benchmark. In *THE WEB CONFERENCE 2025*.

[274] Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, and Chengkai Li. 2025. LLMTaxo: Leveraging Large Language Models for Constructing Taxonomy of Factual Claims from Social Media. In *Findings of the Association for Computational Linguistics: ACL 2025*. 19627–19641.

[275] Jinghui Zhang, Kaiyang Wan, Longwei Xu, Ao Li, Zongfang Liu, and Xiuying Chen. 2025. From Individuals to Crowds: Dual-level Public Response Prediction in Social Media. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25)*. 5903–5912.

[276] Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 14544–14607.

[277] Kexun Zhang, Jane Dwivedi-Yu, Zhaojiang Lin, Yuning Mao, William Yang Wang, Lei Li, and Yi-Chia Wang. 2025. Extrapolating to Unknown Opinions Using LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*. 7819–7830.

[278] Linhao Zhang, Li Jin, Xiaoyu Li, Xian Sun, Xin Wang, Zequn Zhang, Jian Liu, Zhicong Lu, and Guangluan Xu. 2025. Flexible Optimal Transport with Contrastive Graphical Modeling for Multimodal Hate Detection. *IEEE Transactions on Multimedia* 27 (2025), 6397–6409.

[279] Linhao Zhang, Li Jin, Xian Sun, Guangluan Xu, Zequn Zhang, Xiaoyu Li, Nayu Liu, Qing Liu, and Shiyao Yan. 2023. TOT: Topology-Aware Optimal Transport for Multimodal Hate Detection. *Proceedings of the AAAI Conference on Artificial Intelligence* 37, 4 (2023), 4884–4892.

[280] Linhao Zhang, Li Jin, Guangluan Xu, Xiaoyu Li, Cai Xu, Kaiwen Wei, Nayu Liu, and Haonan Liu. 2024. CAMEL: Capturing Metaphorical Alignment with Context Disentangling for Multimodal Emotion Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 8 (2024), 9341–9349.

[281] Litian Zhang, Xiaoming Zhang, Chaozhuo Li, Ziyi Zhou, Jiacheng Liu, Feiran Huang, and Xi Zhang. 2024. Mitigating Social Hazards: Early Detection of Fake News via Diffusion-Guided Propagation Path Generation. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24)*. 2842–2851.

[282] Ming Zhang, Shenghan Zhang, Zhenjie Yang, Lekai Chen, Jinliang Zheng, Chao Yang, Chuming Li, Hang Zhou, Yazhe Niu, et al. 2022. GoBigger: A Scalable Platform for Cooperative-Competitive Multi-Agent Interactive Simulation. In *The Eleventh International Conference on Learning Representations*.

[283] Wenyuan Zhang, Tianyun Liu, Mengxiao Song, Xiaodong Li, and Tingwen Liu. 2025. SOTOPIA-\ensuremathØmega: Dynamic Strategy Injection Learning and Social Instruction Following Evaluation for Social Agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 24669–24697.

[284] Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, et al. 2025. *SocioVerse: A World Model for Social Simulation Powered by LLM Agents and a Pool of 10 Million Real-World Users*. arXiv:2504.10157

[285] Xinglang Zhang, Yunyao Zhang, ZeLiang Chen, Junqing Yu, Wei Yang, and Zikai Song. 2026. Logical Phase Transitions: Understanding Collapse in LLM Logical Reasoning. arXiv:2601.02902 [cs.AI]

[286] Yunyao Zhang, Yihao Ai, Zuocheng Ying, Qirui Mi, Junqing Yu, Wei Yang, and Zikai Song. 2026. Coupling Macro Dynamics and Micro States for Long-Horizon Social Simulation. arXiv:2604.05516 [cs.SI]

[287] Yiming Zhang, Sravani Uttara Nanduri, Liwei Jiang, Tongshuang Wu, and Maarten Sap. 2023. BiasX: "Thinking Slow" in Toxic Content Moderation with Explanations of Implied Social Biases. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

[288] Yunyao Zhang, Zikai Song, Hang Zhou, Wenfeng Ren, Yi-Ping Phoebe Chen, Junqing Yu, and Wei Yang. 2025. GA-S3: Comprehensive Social Network Simulation with Group Agents. In *Findings of the Association for Computational Linguistics: ACL 2025*. 8950–8970.

[289] Yichao Zhang, Zejian Wang, Huangxin Zhuang, Lei Song, Guanghui Wen, Jihong Guan, and Shuigeng Zhou. 2025. Predicting Participation Shift of Users at the Next Stage in Social Networks. *IEEE Transactions on Network Science and Engineering* 12, 2 (2025), 1066–1079.

[290] Yunyao Zhang, Zuocheng Ying, Xinglang Zhang, Junqing Yu, Peng Fang, Xu Chen, Wei Yang, and Zikai Song. 2026. IntervenSim: Intervention-Aware Social Network Simulation for Opinion Dynamics. arXiv:2604.06600 [cs.SI]

[291] Yunyao Zhang, Xinglang Zhang, Junxi Sheng, Wenbing Li, Junqing Yu, Yi-Ping Phoebe Chen, Wei Yang, and Zikai Song. 2026. Semantic-Aware Logical Reasoning via a Semiotic Framework. arXiv:2509.24765 [cs.AI]

[292] ZhaoDan Zhang, Zhao Zhang, Jin Zhang, Hui Xu, and Xueqi Cheng. 2025. MPVStance: Mitigating Hallucinations in Stance Detection with Multi-Perspective Verification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1053–1067.

[293] Chenye Zhao, Yingjie Li, Cornelia Caragea, and Yue Zhang. 2024. ZeroStance: Leveraging ChatGPT for Open-Domain Stance Detection via Dataset Generation. In *Findings of the Association for Computational Linguistics ACL 2024*. 13390–13405.

[294] Jin Zhao, Jingxuan Tu, Han Du, and Nianwen Xue. 2024. Media Attitude Detection via Framing Analysis with Events and Their Relations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 17197–17210.

[295] Tanglong Zhao, Ruifang He, Jing Xu, and Bo Wang. 2024. MultiSum: A Multi-Facet Approach for Extractive Social Summarization Utilizing Semantic and Sociological Relationships. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 17 (2024), 19661–19669.

[296] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue Augmentation with Large Language Models for Emotional Support Conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*. 1552–1568.

[297] Chuanxiong Zheng, Lei Zhang, Hui Wang, Randy Gomez, Eric Nichols, and Guangliang Li. 2024. Shaping Social Robot to Play Games with Human Demonstrations and Evaluative Feedback. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. 9517–9523.

[298] Yilun Zheng, Sha Li, Fangkun Wu, Yang Ziyi, Lin Hongchao, Zhichao Hu, Cai Xinjun, Ziming Wang, Jinxuan Chen, et al. 2025. FanChuan: A Multilingual and Graph-Structured Benchmark For Parody Detection and Analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*. 21937–21957.

[299] Hang Zhou, Jiale Cai, Yuteng Ye, Yonghui Feng, Chenxing Gao, Junqing Yu, Zikai Song, and Wei Yang. 2025. Video anomaly detection with motion and appearance guided patch diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 10761–10769.

[300] Jinfeng Zhou, Zhuang Chen, Dazhen Wan, Bosi Wen, Yi Song, Jifan Yu, Yongkang Huang, Pei Ke, Guanqun Bi, et al. 2024. CharacterGLM: Customizing Social Characters with Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*. 1457–1476.

[301] Qinhong Zhou, Hongxin Zhang, Xiangye Lin, Zheyuan Zhang, Yutian Chen, Wenjun Liu, Zunzhe Zhang, Sunli Chen, Lixing Fang, et al. 2025. Virtual Community: An Open World for Humans, Robots, and Society. In *The Fourteenth International Conference on Learning Representations*.

[302] Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, et al. 2023. SOTOPIA: Interactive Evaluation for Social Intelligence in Language Agents. In *The Twelfth International Conference on Learning Representations*.

[303] Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA Frames: Contextual Reasoning about Effects and Harms of Offensive Statements. In *Findings of the Association for Computational Linguistics: ACL 2023*. 6294–6315.