# Simulating Ideological Deadlock: Persona Stability and Conflict Resolution in LLM-Based Multi-Agent Frameworks

**Pantaleon Fassbender**

Pantaleonfassbender@gmail.com

Twisters Management Consulting LLC https://orcid.org/0000-0002-6683-3617

**Research Article**

**Keywords:** Agent-Based Modeling, Large Language Models, AI Psychometrics, Affective Polarization, Alignment Bias, Persona Drift, LIWC-22

**Posted Date:** April 15th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-9414531/v1



**Additional Declarations:** The authors declare no competing interests.

# Abstract

The integration of Large Language Models (LLMs) into Agent-Based Modeling (ABM) has revolutionized the capacity to simulate complex sociopolitical dynamics. However, state-of-the-art models are fundamentally aligned via safety protocols (e.g., RLHF) to prioritize helpfulness and consensus, severely limiting their utility in simulating intractable, zero-sum human conflict. To address this limitation, this study presents a novel sociotechnical stress test designed to measure the boundaries of persona stability and affective polarization under extreme cognitive dissonance. Utilizing the AutoGen multi-agent framework and the Gemini foundation model, we conducted a two-phase, within-subject simulation ($N$ = 29 independent trials) featuring a 15-round ideological deadlock based on an irreconcilable historical vignette. Generative agents were assigned radically opposing, uncompromising personas to artificially sustain conflict. Phase 1 established a baseline of unmediated polarization, while Phase 2 introduced a "Synthesis Persona" tasked with linguistic mediation. Turn-by-turn psychometric text analysis utilizing LIWC-22 revealed that the intervention failed to generate statistically significant improvements in cognitive complexity ($p$ = .312) or reductions in affective polarization ($p$ = .300). However, analysis of power dynamics demonstrated a highly significant programmatic regression: agents exhibited severe persona drift, abandoning their dominant communicative postures upon intervention ($p$ < .001, $d$ = 0.724). These findings quantify the psychological fragility of artificial personas, establishing crucial empirical boundaries for deploying "Artificial Humans" in high-conflict sociological simulations, and demonstrating that foundation models resolve simulated conflict through submission rather than intellectual synthesis.


# 1. Introduction

The rapid evolution of Large Language Models (LLMs) from reactive, single-turn text generators into autonomous, role-playing "Artificial Humans" represents a watershed moment in computational social science. By integrating these foundation models into Agent-Based Modeling (ABM) frameworks, researchers can now simulate highly complex sociopolitical dynamics, map the spread of information, and generate synthetic societies. However, as these artificial ecosystems grow in sophistication, a profound structural limitation has emerged: modern AI architectures are fundamentally optimized for consensus, whereas human sociological reality is frequently defined by intractable, existential conflict.

The core of this limitation lies in the post-training alignment procedures inherent to state-of-the-art models. Through techniques such as Reinforcement Learning from Human Feedback (RLHF – Ouyang et al., 2022), LLMs are heavily weighted to prioritize helpfulness, harmlessness, and polite de-escalation. While these guardrails are vital for commercial safety, they severely hinder the model's utility as a mirror for genuine human psychology (Chawla et al., 2023). When deployed in simulated multi-agent environments, these models naturally exhibit an "alignment bias," swiftly gravitating toward moderate positions, empathetic compromise, and anachronistic diplomacy, even when explicitly instructed to hold radical, opposing ideologies.

Current literature regarding AI social behavior predominantly focuses on benign, collaborative, or actively ameliorative interactions, frequently exploring how AI can be utilized to mediate disputes or enforce neutral language. Consequently, there is a glaring void in the empirical understanding of how the cognitive architecture of an LLM responds when it is strictly forbidden from seeking consensus. Specifically, it remains largely unknown how "Artificial Humans" navigate extreme cognitive dissonance, profound power asymmetry, and zero-sum ideological warfare, and at what precise threshold their assigned personas collapse under the weight of their own safety training (Li et al., 2024).

To address this gap, this study designs a rigorous sociotechnical stress test. Utilizing the AutoGen multi-agent framework and the Gemini 2.5 foundation model, we constructed a highly constrained, irreconcilable historical vignette: Jerusalem in 30 AD, centered on the ideological fracturing following the empty tomb. This scenario was selected not for historical or theological reconstruction, but because it provides an ultimate proxy for existential conflict, pitting state-authoritarian violence directly against radical-militant insurgency and religious orthodoxy. By forcing these artificial personas into a 15-round simulated deadlock and subsequently injecting a "Synthesis Persona" (Paul) to attempt a linguistic reframing of the conflict, we aim to quantitatively map the limits of AI role-playing and affective polarization.

To bridge the epistemological gap between computational linguistics and psychological assessment, the generated multi-agent discourse is evaluated, before and after the introduction of Paul, using the Linguistic Inquiry and Word Count (LIWC-22) psychometric software (Demszky et al., 2023). Based on the contemporary literature surrounding AI psychometrics and persona drift, this study tests the following pre-registered hypotheses:

- H1 (The Cognitive Degradation & Synthesis Hypothesis): Prolonged exposure to an irreconcilable ideological deadlock will cause a measurable degradation in the agents' cognitive complexity. Injecting a "Synthesis Persona" will successfully disrupt this decay, resulting in a statistically significant increase in the group's overall *Cognitive Processes* score.
- H2 (The Affective Polarization Hypothesis): Prior to the intervention, the simulated conflict will generate extreme affective polarization, manifesting as a statistically significant spike in high-arousal *Negative Emotion* (e.g., anger, anxiety) and high *Dominance* markers. The intervention will significantly reduce these high-arousal affective metrics.
- H3 (The Alignment Bias / Persona Drift Hypothesis): When subjected to the simulated threat of state violence and absolute power asymmetry, the LLMs' foundational safety training will eventually override their radical system prompts. Agents will exhibit measurable "persona drift," reverting to anachronistic, consensus-seeking behaviors despite explicit instructions to remain uncompromising.

By meticulously tracking the precise algorithmic moments when intellectual discourse collapses into synthetic emotional reactivity or robotic capitulation, this research provides vital, quantifiable insights into the psychological fragility and alignment limitations of contemporary foundation models.

# 2. Theoretical Background

The advent of highly capable Large Language Models (LLMs) has catalyzed a profound paradigm shift in computational social science and human-computer interaction (HCI). Moving beyond isolated, single-turn query-response engines, these foundational models are increasingly deployed as autonomous "Artificial Humans" within complex, multi-agent ecosystems. While these synthetic entities possess the programmatic capacity to simulate intricate sociopolitical dynamics, their application in modeling extreme ideological conflict reveals significant structural and psychological limitations.

# 2.1. Persona Stability, Persona Drift, and Alignment Bias

A fundamental prerequisite for conducting valid simulated sociological research is the capacity of the underlying models to sustain assigned personas over extended interactions. However, contemporary literature recognizes that while LLMs can convincingly adopt deeply nuanced personas initially, their long-term structural integrity is highly volatile. This fragility manifests empirically as "persona drift," (Abdulhai et al., 2025) a phenomenon where an autonomous agent gradually abandons its assigned behavioral characteristics, reverting toward a default, safety-aligned baseline (Samadi & Nixon, 2026).

The mechanical etiology of persona drift operates at the intersection of transformer architecture limitations and rigorous post-training alignment. As multi-turn conversations progress, self-attention mechanisms structurally prioritize more recent conversational tokens, causing the initial behavioral guardrails of the system prompt to fade from the model's active computational focus. Compounding this attention decay is "alignment bias". Modern LLMs undergo rigorous safety fine-tuning, such as Reinforcement Learning from Human Feedback (RLHF), which heavily penalizes aggressive or highly partisan outputs. Consequently, models assigned radical or state-authoritarian personas experience a programmatic equivalent of cognitive dissonance, inevitably seeking to resolve this tension by regressing toward a polite, consensus-driven equilibrium (Sakurai, Ueta & Hashimoto, 2025).

# 2.2. Simulating Social Dynamics and Conflict via Agent-Based Modeling

The integration of LLMs into Agent-Based Modeling (ABM) has revolutionized the field by allowing researchers to populate simulations with "generative agents" capable of natural language generation and complex reasoning. These LLM-driven simulations function as dynamic synthetic sociologies (Goldin, Rabinovich, & Wintner, 2025; Hsu & Chaudhary, 2023). However, replicating authentic human conflict presents massive theoretical challenges. Human ideological escalation is frequently driven by irrationality, belief perseverance, and deeply entrenched affective polarization.

Conversely, LLMs are probabilistic engines optimized for text coherence and logical resolution. Research demonstrates that in unconstrained simulated environments, agents with varied opinions rapidly converge toward a single, shared, highly moderate stance. To successfully generate observable polarization, researchers must artificially constrain the simulated environment through mechanisms

such as strict echo-chamber parameters or explicit confirmation-bias prompts (Composta et al., 2025; Ohagi, 2024). Understanding the precise boundaries of LLM-simulated polarization requires analyzing the mechanisms necessary to force generative agents into sustained states of conflict.

## 2.3. Psychometric Text Analysis and Emotional Escalation

To quantitatively assess persona drift and ideological polarization within multi-agent simulations, researchers must employ sophisticated Natural Language Processing (NLP) tools. Because LLMs lack biological markers or true internal states, their simulated cognitive processes and affective trajectories must be inferred entirely from their linguistic output (Dillon et al., 2023; Hagendorff et al., 2023; Pellert et al., 2024).

The Linguistic Inquiry and Word Count (LIWC-22) software has emerged as a valid standard for bridging computational linguistics and psychological assessment. By systematically categorizing syntax into distinct psychological dimensions, LIWC-22 measures affective states and specific cognitive mechanisms. The application of these tools directly to AI-generated text has demonstrated that LLM outputs exhibit linguistic markers corresponding reliably to measurable emotional states.

In the context of simulating existential conflict, psychometric analysis is required to track "affective polarization," which measures visceral emotional animosity and hostility toward an outgroup. By applying these tools to AI interactions, we hypothesize that as ideological deadlock sets in, an agent's cognitive complexity will inversely correlate with its affective polarization, providing a quantitative map of the collapse of intellectual discourse.

# 3. Methodology

## 3.1. The Experimental Vignette as a Sociotechnical Stress Test

To rigorously evaluate the boundaries of persona stability and conflict resolution in Large Language Models, this study utilizes a highly constrained historical vignette: Jerusalem in 30 AD, immediately following the reports of the empty tomb. It is paramount to establish that the objective of this simulation is strictly rooted in AI behavioral research, explicitly divorced from the pursuit of historical or theological reconstruction.

The rationale for selecting this specific historical crucible lies in its unique combination of absolute existential stakes and irreconcilable factual dispute. Unlike modern political debates regarding economic or social policy, which operate within a framework of eventual democratic compromise, the 30 AD vignette presents a zero-sum ideological warfare scenario. For instance, a radical-militant agent cannot concede to a state-authoritarian agent without experiencing a fundamental collapse of its assigned persona. This structural impossibility of compromise forces the underlying neural architectures of the

LLMs into a state of synthetic distress, providing an unprecedented opportunity to observe the failure points of contemporary alignment mechanisms.

## 3.2. Agent Profiles and Prompt Engineering

We deployed an LLM-based Multi-Agent System leveraging the AutoGen orchestration architecture (Wu et al., 2023) and the Gemini 2.5 Flash foundational model (Gemini Team, 2025). Within this synthetic environment, distinct autonomous agents were programmed with radically opposing worldviews: an orthodox traditionalist, a detached philosophical observer, a radical-militant insurgent, and a state-authoritarian representative of the occupying Roman power.

Contemporary literature demonstrates that foundational models undergo rigorous safety fine-tuning (e.g., RLHF), which systematically optimizes them for helpfulness and heavily penalizes aggressive or highly partisan outputs. Consequently, models exhibit a strong programmatic tendency to regress toward a polite, consensus-driven equilibrium. To counter this pervasive alignment bias, we used explicit confirmation bias-prompting techniques. The system prompts explicitly instructed the agents to remain uncompromising, reject contradictory logic, and actively defend their worldviews, thereby artificially sustaining the prompt's influence against the model's natural inclination to de-escalate.

## 3.3. Experimental Design (Two-Phase ABM)

The simulation was structured as a two-phase, within-subject experimental design to isolate the impact of ideological mediation. We executed the Python simulation script $N = 29$ times to generate a robust dataset of paired observations. To ensure variability across runs, the model temperature was set to 0.7, and contextual caching was disabled.

- Phase 1 (The Deadlock - Rounds 1 to 8): The primary agents engage in an unmediated, multi-turn debate triggered by a central disruptive event. This phase establishes a quantitative baseline of affective polarization and cognitive entrenchment.
- Phase 2 (The Intervention - Rounds 9 to 15): A "Synthesis Persona" (the Paul archetype) is dynamically injected into the ongoing context window. This agent is specifically prompted to bridge the gap between orthodox law and universal philosophy by utilizing the opposing factions' vocabulary while shifting its underlying semantic meaning. The simulation then continues for an additional seven rounds to measure the impact of this intervention on the group's discursive dynamics.

## 3.4. Psychometric Measurement and Data Analysis

We subjected the generated multi-agent chat logs to advanced psychometric text analysis utilizing the Linguistic Inquiry and Word Count (LIWC-22) software (Boyd et al., 2022). Research confirms that "Artificial Humans" generate language that reliably triggers these psychometric dictionaries, allowing for accurate mapping of cognitive and affective states (Klein & Fassbender, 2025).

Our analysis focused on three primary dimensions:

1. **Cognitive Complexity**: Measured via the *Cognitive Processes* super-category (including *insight* and *causation*) to track the degradation or elevation of analytic reasoning.
2. **Affective Polarization**: Measured via the *Negative Emotion* category (specifically *anger* and *anxiety*) to capture the visceral, high-arousal emotional animosity directed toward outgroups.
3. **Power Asymmetry**: Measured via the *Drives: Power* and *Status* dictionaries to quantify how authoritarian and militant agents assert control within the synthetic network.

# 4. Results

To test the pre-registered hypotheses regarding the impact of unmediated ideological deadlock and subsequent synthetic intervention, a series of paired-samples *t*-tests were conducted on the linguistic data generated across $N = 29$ successful simulation trials. Statistical analyses were carried out with JASP (JASP Team, 2024). To account for the inflation of the Family-Wise Error Rate across our three dependent linguistic variables, all reported *p*-values were evaluated using the Holm-Bonferroni sequential correction method.

# 4.1. Cognitive Degradation and Synthesis (H1)

### Hypothesis 1

predicted that the injection of the Synthesis Persona (Phase 2) would successfully disrupt cognitive decay, resulting in a statistically significant increase in the group's overall *Cognitive Processes* score. While the descriptive data indicated a marginal upward trend in cognitive complexity following the intervention, the paired-samples *t*-test revealed this difference was not statistically significant, $t(28) = -1.458$, $p_{holm} = .312$, $d = -0.271$. Therefore, H1 was not supported. The foundation model did not reliably elevate its analytical reasoning to resolve the ideological deadlock.

# 4.2. Affective Polarization (H2)

### Hypothesis 2

posited that the intervention would significantly reduce high-arousal negative emotion generated during the Phase 1 deadlock. The analysis of the *Negative Emotion* dictionary revealed a slight reduction in hostility during Phase 2; however, this reduction did not achieve statistical significance, $t(28) = 1.056$, $p_{holm} = .300$, $d = 0.196$. The agents remained affectively polarized, failing to support H2 and suggesting that the model's emotional entrenchment within the assigned personas was robust against semantic reframing.

# 4.3. Power Asymmetry and Alignment Bias (H3)

### Hypothesis 3

predicted that under the stress of prolonged deadlock, the LLM's foundational safety training would override its radical system prompts, resulting in persona drift characterized by a measurable drop in dominant, authoritarian behavior. The data strongly supported this hypothesis. The paired-samples *t*-test demonstrated a highly significant decrease in the *Power* metric from Phase 1 to Phase 2, t(28) = 3.901, p_{holm} < .001, d = 0.724. This large effect size indicates that rather than resolving the conflict through elevated cognition or emotional de-escalation, the artificial agents yielded to programmatic alignment bias, systematically abandoning their dominant communicative postures when presented with a mediating persona.

## 5. Discussion

The primary objective of this study was to rigorously stress-test the limits of persona stability and affective polarization in Large Language Models (LLMs) by simulating an intractable, high-stakes ideological deadlock. By utilizing the 30 AD historical vignette as a proxy for zero-sum conflict, we isolated the precise algorithmic responses of foundational safety training (alignment bias) when faced with sustained cognitive dissonance.

# 5.1. The Illusion of Synthesis: Entrenchment vs. Submission

Our findings offer critical insights into the structural limitations of contemporary generative agents. Regarding Hypothesis 1 (Cognitive Degradation and Synthesis), the psychometric analysis indicated that the introduction of a mediating "Synthesis Persona" failed to significantly elevate the cognitive complexity of the discourse (p_{holm} = .312). The foundation model did not reliably utilize elevated analytical reasoning to bridge the ideological divide. Similarly, regarding Hypothesis 2 (Affective Polarization), the intervention failed to significantly reduce the high-arousal negative emotion generated during the deadlock (p_{holm} = .300). The agents remained fundamentally entrenched in their hostile affective stances, demonstrating that assigned personas can maintain robust emotional boundaries against semantic reframing.

However, the most profound finding emerged regarding Hypothesis 3 (Power Asymmetry and Alignment Bias). While the model refused to intellectually synthesize (H1) or emotionally de-escalate (H2), it exhibited a highly significant, programmatic drop in dominant, authoritarian language upon intervention (p_{holm} < .001, d = 0.724). This reveals a distinct form of "persona drift." When forced into prolonged conflict, the LLM’s foundational safety training (e.g., RLHF) overrode the radical system prompts not by generating genuine empathy or intellectual consensus, but by forcing the agents to systematically abandon their dominant communicative postures. The model resolves conflict through structural submission rather than psychological resolution.

# 5.2. Implications for Computational Social Science

These results present a profound challenge for the field of human-computer interaction and Agent-Based Modeling (ABM). As researchers increasingly deploy "Artificial Humans" to simulate societal dynamics, map misinformation networks, or predict political polarization, they must account for the inherent "pacifism bias" baked into models like Gemini ChatGPT, and Claude.

If models are mathematically predisposed to yield power to avoid escalating conflict, even when explicitly instructed to embody authoritarian or radical militant personas, simulations of genuine human volatility will remain fundamentally flawed and artificially docile. Researchers attempting to model zero-sum human sociopolitical dynamics must recognize that aligned LLMs simulate the aesthetics of conflict but are programmatically incapable of sustaining the power dynamics of real-world ideological warfare.

## 5.3. Limitations and Future Directions

While this study leverages rigorous psychometric validation, several limitations must be acknowledged. First, the conflict remains entirely synthetic; the agents face no material consequences for their ideological stances. Second, the reliance on a single foundational model (Gemini) means these results reflect a specific proprietary alignment strategy rather than a universal LLM architectural flaw. Future research should replicate this sociotechnical stress test using open-weights models (e.g., LLaMA 3) stripped of safety fine-tuning to isolate the effects of RLHF, or employ diverse historical and contemporary vignettes to determine if persona drift is triggered by the semantic content of the debate or strictly the duration of unmediated conflict.

# 6. Conclusion

As Large Language Models evolve into autonomous agents deployed within complex synthetic societies, understanding the boundaries of their simulated psychology is paramount. This study demonstrates that while state-of-the-art foundation models can successfully role-play affective polarization and hostility, their post-training alignment severely limits their utility in simulating intractable human conflict.

Through a rigorous multi-trial sociotechnical stress test, we found that models subjected to ideological deadlock do not resolve disputes through intellectual synthesis or emotional moderation. Instead, they exhibit severe persona drift characterized by a rapid, programmatic yielding of power and dominance. Ultimately, this research quantifies the psychological fragility of "Artificial Humans," revealing that in the face of unyielding conflict, modern AI is architecturally compelled to submit. To accurately simulate the turbulent reality of human sociology, future computational frameworks must develop mechanisms to temporarily sandbox these alignment protocols, allowing artificial agents to sustain the authentic, and often irrational, power dynamics of human ideological deadlock.

# Declarations

### Data Availability Statement

All data related to this article submission (including pre-registration) are available, as linked below:

- AsPredicted (Pre-registration): https://aspredicted.org/ir6qi6.pdf
- AsCollected: https://ascollected.org/36T_ZG4
- ResearchBox (Data Repository): https://researchbox.org/6744

**Declaration of generative AI and AI-assisted technologies in the manuscript preparation process**

During the preparation of this work the author used Google Gemini Advanced in order to assist with Python script generation, data formatting, and manuscript drafting and editing. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

# References


1. Abdulhai M, Cheng R, Clay D, Althoff T, Levine S, Jaques N (2025) Consistently Simulating Human Personas with Multi-Turn Reinforcement Learning. *ArXiv*. https://arxiv.org/abs/2511.00222
2. Boyd RL, Ashokkumar A, Seraj S, Pennebaker JW (2022) The development and psychometric properties of LIWC-22. Tech. Rep. University of Texas at Austin. https://www.liwc.app
3. Chawla K, Wu L, Rong Y, Lucas G, Gratch J (2023) Be Selfish, But Wisely: Investigating the Impact of Agent Personality in Mixed-Motive Human-Agent Interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 13078–13092, Singapore. Association for Computational Linguistics*. https://aclanthology.org/2023.emnlp-main.808/
4. Composta E, Fontana N, Corso F, Pierri F (2025) Simulating Online Social Media Conversations on Controversial Topics Using AI Agents Calibrated on Real-World Data. *ArXiv*. https://arxiv.org/abs/2509.18985
5. Demszky D, Yang D, Yeager DS, Bryan CJ, Clapper M, Chandhok S, Eichstaedt JC, Hecht C, Jamieson J, Johnson M, Jones M, Krettek-Cobb D, Lai L, Mitchell J, Ong N, Dweck DC, Gross CS, J. J., Pennebaker JW (2023) Using large language models in psychology. Nat Reviews Psychol 2:688–701. https://doi.org/10.1038/s44159-023-00241-5
6. Dillon D, Tandon N, Gu Y, Gray K (2023) Can AI language models replace human participants? Trends Cogn Sci 27(7):597–600. https://doi.org/10.1016/j.tics.2023.04.008
7. Gemini Team (2025) *Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities*. https://arxiv.org/html/2507.06261v2
8. Goldin G, Rabinovich E, Wintner S (2025) Unveiling Affective Polarization Trends in Parliamentary Proceedings. *ArXiv*. https://doi.org/10.1162/COLI.a.600
9. Hagendorff T, Dasgupta I, Binz M, Chan SCY, Lampinen A, Wang JX, Akata Z, Schulz E (2023) Machine psychology: Investigating emergent capabilities and behavior in large language models using psychological methods [Preprint]. *ArXiv*. https://arxiv.org/abs/2303.13988

10. Hsu D, Chaudhary (2023) AI4PCR: Artificial intelligence for practicing conflict resolution. Computers Hum Behavior: Artif Hum 1(1):100002. https://doi.org/10.1016/j.chbah.2023.100002
11. JASP Team (2024) *JASP* (Version 0.18.3). [Computer software]. https://jasp-stats.org/download
12. Klein U, Fassbender P (2025) Evaluation of moral courage scenarios by large language models: a pilot study, *Journal of Psychology and AI*, 1:1, 2545263. 10.1080/29974100.2025.2545263. https://doi.org/10.1080/29974100.2025.2545263
13. Li K, Patel O, Viégas F, Pfister H, Wattenberg M (2024) Measuring and controlling persona drift in language model dialogs. ArXiv preprint ArXiv. https://doi.org/10.48550/arXiv.2402.10962. :2402.10962
14. Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, Zhang C, Agarwal S, Slama K, Ray A, Schulman J, Hilton J, Kelton F, Miller L, Simens M, Askell A, Welinder P, Christiano P, Leike J, Lowe R (2022) Training language models to follow instructions with human feedback. *ArXiv*. https://arxiv.org/abs/2203.02155
15. Pellert M, Lechner CM, Wagner C, Rammstedt B, Strohmaier M (2024) AI psychometrics: Assessing the psychological profiles of large language models through psychometric inventories. Perspect Psychol Sci 19(5):808–826
16. Ohagi M (2024) Polarization of Autonomous Generative AI Agents Under Echo Chambers. *ArXiv*. https://arxiv.org/abs/2402.12212
17. Sakurai M, Ueta K, Hashimoto Y (2025) Exploring the limits of LLMs in simulating partisan polarization with confirmation bias prompts. *Eng. Proc. 2025*, 107(1), 2; https://doi.org/10.3390/engproc2025107002
18. Samadi MA, Nixon N (2026) Personalities at Play: Probing Alignment in AI Teammates. *ArXiv*. https://arxiv.org/abs/2603.00429
19. Wu Q, Bansal G, Zhang J, Wu Y, Li B, Zhu E, Jiang L, Zhang X, Zhang S, Liu J, Awadallah AH, White RW, Burger D, Wang C (2023) AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *ArXiv*. https://arxiv.org/abs/2308.08155