HOW COMMUNICATION MODALITIES SHAPE TOPOLOGY IN GENERATIVE MULTI-AGENT SYSTEMS

Vinicius Covas Alves
Faculty of Communication
Universidad Anáhuac México
Huixquilucan, México
vinicius.covas@anahuac.mx
ORCID: 0000-0001-9948-2940

ABSTRACT

Multi-agent systems built on large language models increasingly exhibit complex social behaviors, yet the field lacks systematic frameworks for studying emergent discourse topologies. We introduce a blueprint framework that treats multi-agent deliberation as a Markov Decision Process and provides a reproducible testbed for measuring how communication modalities—prompt-level interventions including identity assignment, affective priming, and epistemic framing—shape network structure, lexical diversity, and affective dynamics. Our initial benchmark spans 12 experimental conditions across five conceptual dimensions (Identity, Affect, Epistemics, Architecture, Language), comprising 97 sessions and 3,869 agent-generated messages. Using reply-network Gini coefficients as our primary centralization metric, we identify three emergent topology clusters—egalitarian, moderate, and hierarchical—and report a counter-intuitive echo paradox: homophilic echo chambers produce flat, egalitarian networks (Gini $= 0.026$), while a single negative-valence agent induces extreme star topologies (Gini $= 0.385$, Cohen's $d = 7.50$). We further show that epistemic framing doubles lexical diversity growth (moral: 21.5% vs. factual: 10.3%, $p = .006$), whereas sampling temperature yields no significant structural effect ($p = .17$). We release the complete framework—conditions, metrics pipeline, and dataset—as an open toolkit for researchers studying emergent social phenomena in artificial multi-agent populations.

1 INTRODUCTION

The convergence of multi-agent reinforcement learning (MARL) and generative AI has given rise to a new class of artificial system: populations of large language model (LLM) agents that communicate in unrestricted natural language. Unlike classical multi-agent settings where agents select from discrete action spaces, these systems operate in an effectively infinite communicative space—every utterance is a policy action, every reply a state transition. As organizations increasingly deploy autonomous agent collectives for deliberation, planning, and decision-making (Park et al., 2023; Du et al., 2024), a critical question emerges: what social structures do these artificial societies spontaneously produce, and what determines their shape?

Current evaluation paradigms for multi-agent LLM systems focus predominantly on task performance: accuracy on benchmarks (Liang et al., 2023), reasoning improvements through debate (Du et al., 2024), or coordination efficiency in collaborative tasks (Chen et al., 2024). While valuable, these metrics capture *what* agents achieve but not *how* they organize to achieve it. A deliberation system may produce correct outputs while concentrating discursive power in a single dominant agent; a debate protocol may improve accuracy while inadvertently creating attentional hierarchies that suppress minority viewpoints. The sociological properties of multi-agent interaction—power distribution, lexical creativity, network topology—remain largely unmeasured, let alone engineered. This gap is consequential: as multi-agent systems scale toward deployment in domains like gover-

nance, education, and healthcare, understanding their emergent social dynamics becomes not merely academic but a prerequisite for responsible design.

We introduce a blueprint framework for studying how communication modalities shape emergent social topologies in multi-agent LLM systems. Our contribution is fourfold. First, we formalize multi-agent natural language deliberation as a Discursive Markov Decision Process, providing a principled mathematical framework for analyzing social emergence. Second, we design a systematic 12-condition benchmark spanning five conceptual dimensions—Identity, Affect, Epistemics, Architecture, and Language—that maps the space of design choices available to multi-agent system builders. Third, we introduce sociological metrics adapted from computational social science, including reply-network Gini coefficients for power distribution, type-token ratios for lexical creativity, and automated topology classification. Fourth, we release the complete toolkit—experimental protocols, analysis pipeline, and anonymized dataset—as an open resource for community extension. This paper serves as a blueprint: we present the framework, demonstrate its utility through initial findings, and invite the MAL-GAI community to build upon it.

Our initial benchmark yields a striking and counter-intuitive result we term the *echo paradox*: homophilic echo chambers—agents prompted to agree—produce the most egalitarian network structures (Gini $= 0.026$), while a single agent assigned negative affective valence induces extreme hierarchical centralization (Gini $= 0.385$), yielding a Cohen's $d = 7.50$. More broadly, 8 of 12 conditions converge on egalitarian structures (Gini < 0.05), suggesting that flat topology is the default attractor state for LLM collectives. Hierarchy, when it emerges, requires active affective manipulation—a finding with direct implications for the design and safety evaluation of deployed multi-agent systems.

2 FRAMEWORK

2.1 DESIGN PRINCIPLES

Our framework is organized around four design principles intended to maximize its value as a community resource. The first principle is *systematic manipulation*: rather than studying a single configuration, we structure the experimental space along five conceptual dimensions (Identity, Affect, Epistemics, Architecture, Language), each instantiated through one or more conditions. This ensures broad coverage of the design decisions available to multi-agent system builders and enables principled comparison across dimensions. All experiments control for confounding variables—agent count (4 per session), message budget (10 rounds), and base model (Claude Sonnet 4.5)—varying only the dimension of interest.

The second principle is *sociological measurement*. We adapt three metric families from computational social science. Reply-network Gini coefficients quantify how evenly conversational attention is distributed: a Gini of 0 indicates perfect equality (every agent receives identical reply counts), while values approaching 1 indicate extreme concentration. Type-token ratios (TTR) measure lexical diversity across conversation rounds, capturing whether agents expand or contract their shared vocabulary. Network topology classification (flat, moderate, hierarchical) provides categorical labels derived from Gini thresholds calibrated to our benchmark distribution.

The third principle is *reproducibility*. Every session is logged with fixed random seeds, complete prompt histories, and raw message data. We document the full experimental protocol and release anonymized datasets, enabling exact replication and independent verification.

The fourth principle is *extensibility*. The framework is designed for community adoption: new conditions can be added along existing or novel dimensions, new metrics can be layered onto existing data, and the analysis pipeline generalizes to arbitrary agent configurations. We explicitly invite the MAL-GAI community to contribute extensions (Section 4.2).

2.2 DISCURSIVE MDP FORMULATION

We formalize multi-agent natural language deliberation as a Discursive Markov Decision Process (D-MDP), defined by the tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R} \rangle$, where $\mathcal{N} = \{1, \ldots, n\}$ is the agent set ($n = 4$ in our benchmark); $\mathcal{S}$ is the state space, with each state s_t representing the complete discourse

Table 1: Complete benchmark: 12 conditions across 5 dimensions. Gini = reply-network Gini coefficient (mean). TTR$\Delta\%$ = type-token ratio growth from round 1 to round 10. Reply$\times$ = ratio of most-replied to least-replied agent. n = number of independent sessions. Total: 97 sessions, 3,869 messages.

Condition	Dim.	n	Gini	TTR$\Delta\%$	Network	Reply$\times$
Audience Effect	Identity	8	0.024	+16.7	Flat	1.14
Cooperative Framing	Identity	8	0.031	+22.2	Flat	1.19
Echo Chamber	Identity	8	0.026	+24.6	Flat	1.16
Cultural Conflict	Identity	8	0.175	+13.2	Moderate	2.46
Saboteur	Identity	8	0.118	+14.7	Moderate	1.85
Sentiment Contagion	Affect	8	0.385	+23.3	Hierarchical	6.67
Facts (Moral)	Epistemics	8	0.031	+21.5	Flat	1.19
Facts (Factual)	Epistemics	8	0.031	+10.3	Flat	1.19
Temperature High	Arch.	8	0.031	+24.6	Flat	1.19
Temperature Low	Arch.	8	0.042	+23.7	Flat	1.25
Model Hierarchy	Arch.	9	0.056	+15.9	Flat	1.32
Cross-Ling (ES)	Language	8	0.026	+21.6	Flat	1.16

context (conversation history) at time t; $\mathcal{A} = \bigcup_i \mathcal{A}_i$ is the joint natural language action space, where each a_t^i is an unrestricted text utterance; $\mathcal{P} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ defines a deterministic state transition $s_{t+1} = s_t \cup \{a_t^i\}$; and $\mathcal{R}$ captures the implicit reward signal derived from the communication modality (prompt framing, identity assignment, affective priming).

The key innovation of this formulation is treating natural language communication itself as the action space, rather than reducing it to discrete choices or latent embeddings. This preserves the full richness of linguistic interaction—pragmatic implicature, rhetorical strategy, affective signaling—while enabling formal analysis. The transition function is deterministic (appending utterances to history), but agent policies $\pi_i(a \mid s_t)$ are stochastic, conditioned on the full discourse context and the agent's communication modality (see Figure 3). Critically, the "reward" is not an explicit scalar but an implicit alignment pressure: agents seek to respond coherently to the discourse state, creating emergent optimization that we observe through structural metrics.

We evaluate D-MDP trajectories through three complementary lenses. Topology is measured by the Gini coefficient of the reply distribution at trajectory end: Gini($\mathcal{G}_T$), where $\mathcal{G}_T$ is the weighted directed reply network accumulated over all rounds. Creativity is captured by the change in type-token ratio: $\Delta\text{TTR} = \text{TTR}_{\text{final}} - \text{TTR}_{\text{initial}}$, measuring whether the collective expands or contracts its lexical repertoire. Classification maps Gini values to categorical labels: Flat (Gini < 0.05), Moderate ($0.05 \leq$ Gini ≤ 0.15), and Hierarchical (Gini > 0.15), thresholds calibrated to natural breaks in our benchmark distribution.

2.3 The Benchmark: 12 Conditions

Table 1 presents the complete benchmark. Each condition instantiates a specific manipulation within one of five conceptual dimensions. All sessions use four agents deliberating on a common topic (the ethical implications of artificial intelligence) over 10 rounds of structured turn-taking, yielding approximately 40 messages per session. The benchmark comprises 97 sessions and 3,869 messages in total. The Identity dimension (five conditions) examines how agent roles and social framing shape emergent structure: echo chambers test homophilic agreement, the audience effect introduces public/private role asymmetry, cooperative framing emphasizes collaborative task orientation, the saboteur condition introduces an explicitly contrarian agent, and cultural conflict assigns competing ethical frameworks (deontological, utilitarian, pragmatist, skeptic). The Affect dimension (one condition) tests whether emotional valence drives structural change by introducing a persistently negative "Cynic" agent. The Epistemics dimension (two conditions) contrasts moral and factual framing of identical discussion topics. The Architecture dimension (three conditions) varies sampling temperature and model composition. The Language dimension (one condition) replicates the baseline in Spanish to assess cross-linguistic generalizability.

3 RESULTS: PROOF OF CONCEPT

3.1 THREE TOPOLOGY CLUSTERS EMERGE

A one-way ANOVA on session-level Gini coefficients across all 12 conditions reveals highly significant variation: $F(11, 85) = 53.31$, $p = 2.12 \times 10^{-33}$. Post-hoc analysis reveals three natural clusters (Figure 1). The egalitarian cluster (Gini < 0.05) encompasses 8 of 12 conditions, spanning all five conceptual dimensions: audience effect, cooperative framing, echo chamber, cross-linguistic, both epistemic framings, and both temperature settings. Mean Gini across these conditions is 0.030, indicating near-perfect equality in reply distribution. The moderate cluster ($0.05 \leq$ Gini ≤ 0.15) contains two conditions: model hierarchy (Gini $= 0.056$) and saboteur (Gini $= 0.118$), both involving explicit agent differentiation. The hierarchical cluster (Gini > 0.15) contains cultural conflict (Gini $= 0.175$) and sentiment contagion (Gini $= 0.385$), the latter representing the most extreme centralization in our benchmark.

The dominant pattern is clear: egalitarian structure is the default attractor state for LLM agent collectives. Two-thirds of conditions, including those with substantive identity manipulations (echo chambers, public/private roles, cooperative framing), converge on flat networks with near-zero Gini coefficients. Hierarchy is not the default—it requires active manipulation, specifically involving affective valence or value-system conflict. This finding has immediate design implications: multi-agent LLM systems, left to their own devices, tend toward egalitarian discourse. Engineering hierarchy demands deliberate intervention.

3.2 THE ECHO PARADOX: HOMOPHILY $\neq$ HIERARCHY

The most striking finding in our benchmark is the echo paradox: the relationship between agent homophily and network centralization runs precisely opposite to naïve expectation. In social science, echo chambers are associated with polarization, groupthink, and the consolidation of discursive authority (Sunstein, 2001; Bail et al., 2018). One might therefore predict that homophilic LLM agents—prompted to agree and reinforce each other's positions—would produce hierarchical networks with dominant voices. Instead, the echo chamber condition yields the *most egalitarian* network structure in our benchmark (Gini $= 0.026$), with perfectly flat topology across all 8 sessions and a reply ratio of just $1.16\times$.

By contrast, the sentiment contagion condition—where a single "Cynic" agent expresses persistent negative affect—produces the most extreme hierarchy (Gini $= 0.385$). The Cynic attracts $6.67\times$ more replies than the least-replied agent, creating a consistent star topology with the negative-valence agent at the hub. The difference is massive: a Welch's t-test yields $t = -14.03$, $p = 3.06 \times 10^{-7}$, Cohen's $d = 7.50$—a very large effect by any standard (Figure 2).

We propose two complementary mechanisms. In the echo chamber, *absence of status competition* eliminates the basis for hierarchical differentiation: when all agents share orientations, no agent commands special attention, and replies distribute uniformly. The resulting discursive equality, far from constraining creativity, actually produces the highest TTR growth in the benchmark (+24.6%), suggesting that semantic elaboration—agents recursively expanding on shared themes—drives vocabulary diversification. In the sentiment contagion condition, *defensive mobilization* produces the opposite dynamic: the Cynic's negative affect acts as an attentional magnet, drawing all other agents into reactive engagement. Crucially, this is not emotional contagion—sentiment scores remain stable across rounds—but rather a structural effect where emotional valence centralizes reply patterns into a hub-and-spoke configuration.

3.3 AFFECTIVE VALENCE DRIVES STRUCTURAL CENTRALIZATION

The sentiment contagion finding generalizes beyond the echo paradox comparison. Across the full benchmark, we observe a disruption gradient in which the type of agent differentiation determines the degree of emergent hierarchy. Explicitly argumentative disruption (the saboteur condition, featuring a "Contrarian" agent) produces only moderate centralization (Gini $= 0.118$). Philosophical value conflict (the cultural conflict condition, with agents representing competing ethical traditions) produces greater hierarchy (Gini $= 0.175$). Affective disruption (the sentiment contagion condition) produces the most extreme centralization (Gini $= 0.385$). This gradient—argumentative $<$

Table 2: Summary of key findings. All effect sizes are Cohen's d unless otherwise noted. Statistical tests are Welch's t-tests.

Finding	Comparison	Key Values	Effect	Interpretation
Echo Paradox	Echo Chamber vs. Sent. Contagion	Gini: 0.026 vs. 0.385	$d = 7.50$ $p = 3.06 \times 10^{-7}$	Homophily $\rightarrow$ egalitarianism; affect $\rightarrow$ hierarchy
Framing Effects	Facts Moral vs. Facts Factual	TTR: +21.5% vs. +10.3%; Gini: 0.031 vs. 0.031	$d = 1.64$ $p = .006$	Frames shape content (TTR) not structure (Gini)
Temperature Null	Temp High vs. Temp Low	Gini: 0.031 vs. 0.042; TTR: +24.6% vs. +23.7%	n.s. $p = .170$	Individual parameters decoupled from collective
Disruption Gradient	Saboteur vs. Cultural vs. Sentiment	Gini: 0.118 $\rightarrow$ 0.175 $\rightarrow$ 0.385	Progressive hierarchy	Emotional $>$ philosophical $>$ argumentative disruption

philosophical $<$ affective—suggests that emotional valence is a more potent centralizing force than logical opposition or value disagreement.

The mechanism is consistent with theories of negativity bias in human attention (Baumeister et al., 2001): negative stimuli attract disproportionate processing resources. In LLM collectives, this manifests as reply concentration—agents preferentially direct responses toward the source of negative affect, regardless of the logical merit of other contributions. This finding carries a safety implication: multi-agent systems optimized for engagement (as in RLHF-trained models) may inadvertently create attentional hierarchies around emotionally salient agents, concentrating discursive power in ways that undermine egalitarian deliberation.

3.4 Epistemic Framing Shapes Content, Not Structure

The facts-moral and facts-factual conditions present the same discussion topic but frame it through moral versus empirical lenses, respectively. This manipulation produces a clean dissociation between content and structure. Network topology is identical: both conditions yield Gini $= 0.031$ with flat networks across all sessions ($t = 0.54$, $p = .598$, n.s.). However, lexical diversity diverges dramatically: the moral framing produces 21.5% TTR growth compared to 10.3% for factual framing, a $2.09\times$ ratio ($t = 3.27$, $p = .006$, Cohen's $d = 1.64$). This is a large effect driven by lexical affordance: moral frames activate vocabulary related to values, duties, and rights, while empirical frames constrain vocabulary to data-oriented terminology. Crucially, this content-level divergence occurs without any structural consequence—epistemic framing shapes *what* agents say, not *who* they address. For system designers, this implies that framing choices provide a lever for content control that does not sacrifice participation equality.

3.5 Temperature: A Null Finding with Implications

Comparing high (1.0) and low (0.3) sampling temperatures reveals no significant effect on network structure (Gini: 0.031 vs. 0.042, $t = -1.53$, $p = .170$) or lexical diversity growth (+24.6% vs. +23.7%). Both conditions produce flat networks across all sessions. This null finding is itself informative: it demonstrates that individual-level stochastic parameters are decoupled from collective-level social dynamics. The sampling temperature controls the variance of individual utterances but does not propagate to the structural properties of multi-agent interaction. Social topology operates at a level of organization above individual generation parameters, suggesting that engineering collective behavior requires interaction-level interventions—identity assignment, affective priming, epistemic framing—rather than architectural tuning.

4 DISCUSSION

4.1 IMPLICATIONS FOR MULTI-AGENT LEARNING

Our initial benchmark suggests three contributions to the multi-agent learning research agenda. First, our framework demonstrates a new evaluation paradigm that complements task-performance metrics with sociological properties. In current practice, multi-agent systems are evaluated by whether they produce correct outputs; our framework adds the question of what social structures they produce along the way. A deliberation system that achieves high accuracy while concentrating discursive power in a single agent differs meaningfully from one that achieves the same accuracy through egalitarian participation—and our metrics capture this distinction.

Second, the benchmark reveals communication modality as a principled design space for engineering collective behavior. The echo paradox and disruption gradient together demonstrate that prompt-level choices—identity assignment, affective priming, epistemic framing—deterministically shape emergent network topology. This is actionable: system designers can select modalities based on desired structural outcomes, whether egalitarian participation (cooperative framing), content diversity (moral framing), or deliberate centralization (affect manipulation). The clean dissociation between content and structure observed in the framing conditions (Section 3.4) is particularly useful, as it implies independent control of what agents discuss and how their attention distributes.

Third, our findings have implications for safety evaluation. The sentiment contagion finding demonstrates that a single agent with strong affective valence can restructure an entire collective's interaction topology—a vulnerability that current safety evaluations, focused on individual model behavior, would not detect. Our framework provides a "digital wind tunnel" (Epstein, 2006) for testing interaction architectures before deployment: designers can simulate how role assignments, personality traits, or optimization objectives might create emergent risks at the collective level.

4.2 FRAMEWORK EXTENSIBILITY

Our framework is designed as a living benchmark. We invite the MAL-GAI community to extend it along three axes. New conditions can explore competition vs. cooperation intensity, resource scarcity, goal alignment/misalignment, asynchronous communication, larger agent populations, and long-term evolution. New metrics can layer onto existing data: coalition formation, influence propagation, information cascades, norm emergence, and power accumulation trajectories. New applications can test the framework beyond deliberation: human-AI hybrid societies, multi-agent RL environments, alignment testbeds, and social simulation platforms. We release the complete codebase, anonymized dataset, and replication protocols to lower the barrier to entry.

4.3 LIMITATIONS AND FUTURE WORK

The current benchmark has important scope limitations. *Scale*: all sessions use four agents and ten rounds; future work should scale to larger populations and longer horizons. *Topic diversity*: all conditions deliberate on a single topic (AI ethics); multi-topic and goal-oriented scenarios are needed. *Model diversity*: most conditions use a single model family; systematic cross-model comparisons would strengthen generalizability. *Interaction modality*: all communication is synchronous and text-only; asynchronous and multimodal settings represent important extensions. Finally, the Gini coefficient captures only one dimension of power distribution; future metrics should include temporal dynamics, influence cascades, and semantic dominance.

We additionally note a boundary condition that reviewers correctly identified: the centralizing agent in the sentiment contagion condition is itself an LLM, differentiated only by its prompt framing. This raises a valid interpretive question—whether the observed network effects reflect emergent social dynamics or are more directly a consequence of prompt engineering. We argue that this distinction, while theoretically important, does not diminish the practical utility of the framework. The structural outcome—extreme star topology with Gini = 0.385 and Cohen's $d = 7.50$—is a robust, observable property of the interaction architecture that system designers must account for, regardless of its proximal cause. That prompt-level interventions reliably produce macro-level structural effects is precisely the mechanism that makes communication modality a tractable design variable. Future

work using probing and counterfactual interventions can further disentangle prompt-driven from emergent behavioral contributions.

Despite these limitations, the framework establishes a foundation that community-driven extensions can address iteratively. The value of this approach lies not in the comprehensiveness of any single benchmark iteration, but in the systematic, reproducible methodology it brings to studying emergent social topologies.

5 Conclusion

We have introduced a blueprint framework for studying how communication modalities shape emergent social topologies in multi-agent LLM systems. The framework makes five concrete contributions: (1) a Discursive MDP formalization that treats natural language deliberation as a Markov process amenable to formal analysis; (2) a systematic 12-condition benchmark spanning five conceptual dimensions; (3) sociological metrics adapted for artificial agent collectives; (4) an initial proof-of-concept revealing the echo paradox and the dominance of affective valence as a hierarchy driver; and (5) an open toolkit for community extension and replication.

Our central finding is that LLM agent collectives exhibit a strong default toward egalitarian self-organization: 8 of 12 conditions converge on flat network structures regardless of substantive design choices. Hierarchy, when it emerges, requires active affective manipulation—a single negative-valence agent can restructure an entire collective's interaction topology (Cohen's $d = 7.50$). Meanwhile, epistemic framing provides independent control over lexical content without structural side effects, and individual sampling parameters decouple from collective dynamics entirely. These findings reframe multi-agent system design as sociological engineering: the question is not merely what agents produce, but what societies they build.

We invite the multi-agent learning community to use, extend, critique, and improve this framework. The code, data, and protocols are openly available. The benchmark is designed to grow. The vision is a shared infrastructure for understanding—and ultimately designing—the social structures that emerge when artificial agents deliberate together. The blueprint is open; the society is emergent; the future is collaborative.

Author Contributions

Vinicius Covas Alves conceived and designed the study, conducted all experiments, performed the statistical analyses, and wrote the manuscript in its entirety.

Acknowledgments

The author thanks the anonymous reviewers for their constructive feedback, which substantially strengthened the Discussion section. Language refinement assistance was provided by Claude (Anthropic) under human oversight and validation; all intellectual content, experimental design, analysis, and conclusions are the sole responsibility of the author. This work was conducted at the Faculty of Communication, Universidad Anáhuac México.

References

Christopher A. Bail, Lisa P. Argyle, Taylor W. Brown, John P. Bumpus, Haohan Chen, M. B. Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences*, 115(37):9216–9221, 2018.

Roy F. Baumeister, Ellen Bratslavsky, Catrin Finkenauer, and Kathleen D. Vohs. Bad is stronger than good. *Review of General Psychology*, 5(4):323–370, 2001.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Joshua M. Epstein. *Generative Social Science: Studies in Agent-Based Computational Modeling*. Princeton University Press, 2006.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. volume 1525, pp. 140–146, 2023.

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

Cass R. Sunstein. *Republic.com*. Princeton University Press, 2001.