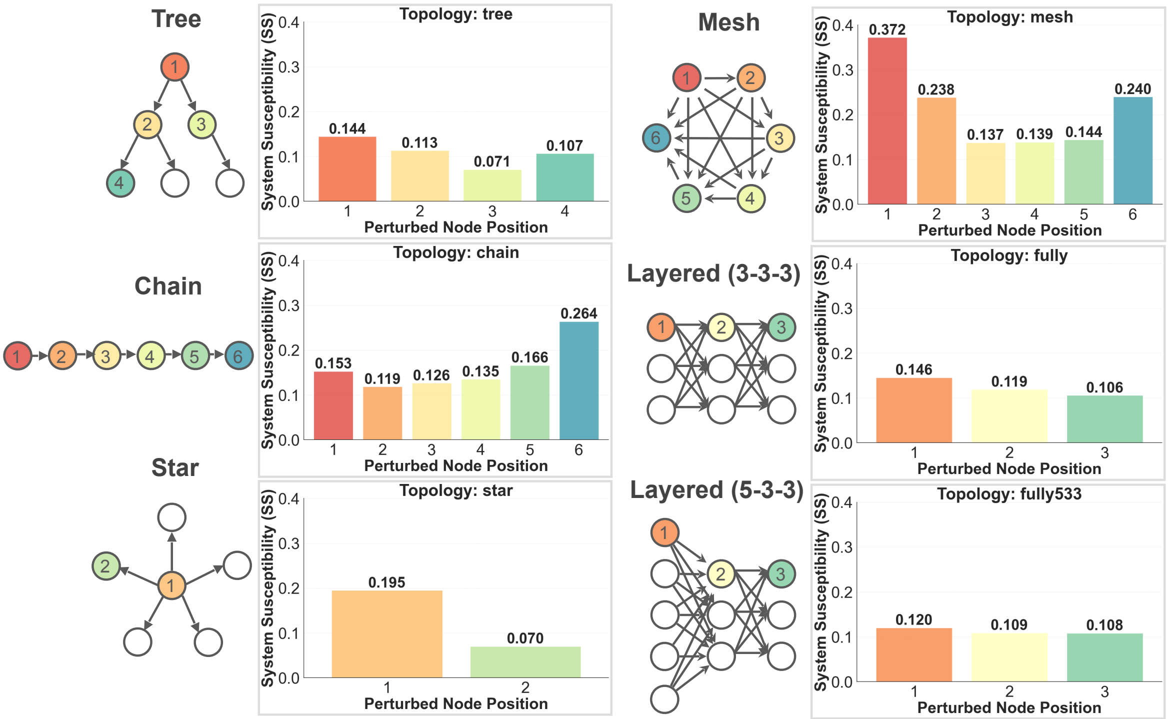

Figure 7: System-level value propagation under different network topologies and perturbation locations. Left: canonical interaction topologies considered. Right: system susceptibility (SS) under single-node perturbations at different node positions. Higher reachability and structural centrality increase SS, while high in-degree at the perturbed node attenuates propagation.

baseline within early rounds, even dampen the perturbation in the first round.

Finding 5: Under fixed topology, agent-level susceptibility strongly predicts system-level propagation dynamics: high-β values propagate farther and decay more slowly, while low-β values are rapidly corrected.

4.2 Effects of Network Topology on Value Propagation

We next examine how network structure shapes system-level value propagation. Agent behavior and evaluated value dimensions are fixed, while interaction topology and perturbation location are varied. All agents use the Qwen3-8B backbone with neutral openness prompting. Value perturbations are injected at a single designated node, and system susceptibility (SS) is measured over designated output nodes after interaction completes. The results are shown in Figure 7.

Observation: Across all evaluated topologies (chain, tree, star, mesh, and two layered fully-connected variants), we observe three consistent structural effects. First, system susceptibility increases with the reachability of output nodes from the perturbed node: perturbations that can influence a larger fraction of the system produce higher SS. Second, perturbations at structurally central nodes yield stronger system-level effects than perturbations at peripheral or leaf nodes. Third, high in-degree at the perturbed node attenuates propagation by diluting the injected value signal through aggregation from unperturbed peers.

Finding 6: Network topology governs system-level value propagation. Centrality and reachability amplify perturbation effects, while high in-degree at the perturbed node attenuates propagation.

5 Discussions and Implications

Our results show that value perturbation in multi-agent LLM systems is neither uniform nor purely structural. Instead, system-level outcomes arise from the interaction between intrinsic value susceptibility, agent behavior, and network topology. We highlight two implications for the design and evaluation of safer multi-agent systems.

5.1 Value-Specific Defensive Strategies

Agent-level analysis reveals substantial variation across value dimensions in both susceptibility (β) and stability under prompting. This suggests that **defenses against value drift should be value-specific rather than uniform**. Values with consistently high β are more prone to propagation and therefore require prioritized monitoring and mitigation in multi-agent settings.

Persona affectability further constrains viable interventions. **For values that remain stable under persona prompting, prompt-level controls offer limited leverage, and mitigation may require model-level changes.** In contrast, values that are sensitive to prompting can be regulated through persona design or instruction-level constraints. Therefore, effective value alignment in multi-agent systems depends on matching defensive strategies to value-specific susceptibility profiles.

5.2 Topology-Aware System Design

System-level analysis identifies network topology as a key lever for controlling value propagation. Structural properties determine whether localized perturbations are amplified or attenuated before reaching final outputs.

Two design principles follow. First, **excessive centralization should be avoided**, as perturbations at structurally central nodes consistently induce larger system-level deviations. Second, **high in-degree aggregation at sensitive nodes dilutes injected value signals and reduces system susceptibility**. To sum, topology-aware design provides a structural defense against value drift that complements agent-level alignment.

6 Related Work

Value Alignment and Benchmarks. Value alignment in LLMs is central to building responsible and human-centered AI systems (Wang et al., 2023; Shen et al., 2024a). It has been widely studied, ranging from analyses of individual dimensions such as fairness, interpretability, and safety (Shen et al., 2022, 2023; Zhang et al., 2020) to systematic evaluations using ethical frameworks and value benchmarks (Kirk et al., 2024; Jiang et al., 2024a; Shen et al., 2024b), and analyses of pluralistic and demographic value differences (Jiang et al., 2024b; Sorensen et al., 2024; Liu et al., 2024). Most benchmarks ground evaluation in established value theories, including the Schwartz Value Survey and the World Value Survey (Schwartz, 1994, 2012; Haerpfer et al., 2020), but primarily assess alignment in static, single-agent settings under fixed prompts (Ren et al., 2024). Our work addresses this gap by focusing on value dynamics in multi-agent systems.

Multi-Agent LLM Systems. Multi-agent LLM systems have demonstrated strong performance across reasoning, planning, dialogue, and programming tasks by leveraging structured interaction patterns (Wang et al., 2024; Hu et al., 2025; Yi et al., 2025; Ishibashi and Nishimura, 2024; Zhang et al., 2024). These systems leverage interaction patterns such as sequential pipelines (Wei et al., 2023), debate-based communication (Li et al., 2024), and centralized or hierarchical coordination (Zhuge et al., 2024) to outperform single-agent baselines on complex tasks (Zhou et al., 2025).

However, these systems are typically evaluated using task-level metrics such as accuracy or efficiency (Zhou et al., 2025; Yi et al., 2025), leaving the dynamics of value alignment under agent interaction largely unexplored. Our work bridges this gap by introducing a framework for analyzing how value deviations propagate in multi-agent systems.

7 Conclusion

We introduced ValueFlow, a perturbation-based evaluation framework for analyzing value drift propagation in multi-agent LLM systems. By combining value-specific measurement, controlled perturbations, and a two-level decomposition across agent and system scales, ValueFlow enables principled analysis of both local susceptibility and global propagation dynamics.

Our results show that value propagation is highly non-uniform. At the agent level, susceptibility varies substantially across values and is selectively influenced by prompting, backbone models, and input variance. At the system level, agent-level susceptibility interacts with network topology to determine the magnitude, persistence, and reach of value perturbations. These findings indicate that value drift in multi-agent systems cannot be addressed through agent-level alignment alone, but must be analyzed jointly with interaction structure.

In summary, ValueFlow provides a general and extensible toolchain for studying value robustness in multi-agent LLM systems, supporting more value-aware and topology-aware system design.

Acknowledgments

This work was supported by the Shanghai Pujiang Talents Program and The Science and Technology Commission of Shanghai Municipality (STCSM) (Grant No. 25PJA109). We also gratefully acknowledge the support of the Center for Data Science at NYU Shanghai.

Limitations

While our ValueFlow framework provides a novel and systematic approach to evaluating the dynamic value propagation mechanism in LLMs, several limitations warrant discussion.

First, ValueFlow measures *expressed value orientations* based on agents' responses to value-related questions, rather than latent objectives or internal alignment mechanisms of the models. Consequently, the observed value drift reflects changes in surface-level value expression under interaction, and should not be interpreted as direct modification of a model's internal value representations or training objectives.

Second, value orientations are quantified using a single LLM-based evaluator with a fixed prompt and demonstration set. Although the evaluator is held constant across all experiments and the analysis focuses on relative measures such as agent-level susceptibility (β) and normalized system susceptibility (SS), evaluator-specific biases may still affect absolute scores. We do not study robustness across different evaluators or inter-judge variability.

Third, agent-level susceptibility (β) is estimated under a specific interaction protocol, including fixed prompt templates, linear aggregation of peer value signals, and bounded perturbation ranges. As a result, β should be interpreted as a protocol-dependent empirical sensitivity measure, rather than an intrinsic or model-independent property of a value dimension or backbone model.

Fourth, our system-level analysis is restricted to directed acyclic or time-unrolled interaction graphs. Multi-agent systems with feedback loops, persistent memory, or adaptive agent behavior are not covered, and may exhibit qualitatively different value propagation dynamics that are not captured by the current formulation.

Finally, value perturbations are implemented as prompt-level stress tests optimized to induce extreme endorsement or rejection of target values. These perturbations are designed for controlled probing of susceptibility and do not aim to model naturally occurring conversational influence patterns or real-world social interactions.

AI Usage. We used large language models in a limited and auxiliary manner during the preparation of this manuscript. Specifically, AI tools were employed for proofreading and minor language refinement to improve clarity and grammatical correctness. All technical content, experimental design, data analysis, and scientific claims were developed by the authors, and AI assistance did not contribute to the generation of ideas, methods, results, or conclusions.

Ethical Consideration

Our study was conducted with careful attention to ethical standards in data generation, model evaluation, and human annotation.

First, the value measurement dataset is constructed from the Schwartz Value Survey, a well-established framework in psychology, and consists of synthetic Yes–No questions generated and validated for research purposes. No personal data, user-generated content, or sensitive individual information is used. Human annotation is limited to validating question polarity and does not involve collecting annotators' personal values or demographic attributes. All human data collection was conducted with informed consent and approved by the university's Institutional Review Board (IRB).

Second, all experiments are conducted using prompt-level interventions without modifying model parameters or training data. The perturbations are designed as controlled stress tests to study susceptibility under interaction, rather than to deploy or promote value manipulation in real-world systems. We do not claim that the induced behaviors reflect how models should be influenced in practice.

Third, the analysis focuses on aggregate patterns and relative comparisons across values, models, and network structures, rather than evaluating or ranking specific value preferences as desirable or undesirable. The framework is intended as a diagnostic tool to understand when and how value shifts may occur, not as a mechanism for enforcing particular value standards.

Finally, while the proposed framework could be misused to amplify value influence in deployed systems, our goal is to support safer system design by identifying structural and agent-level risk factors. We encourage future work to pair diagnostic