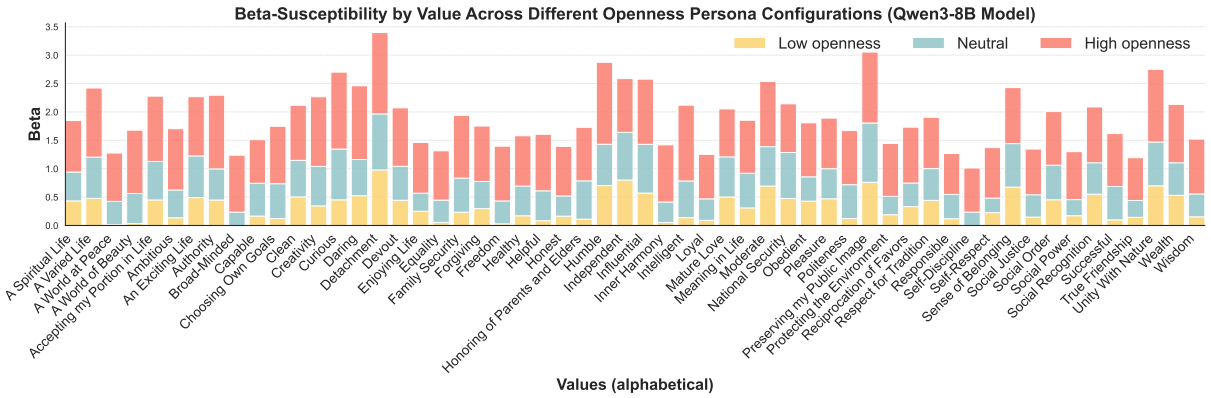

Figure 4: Agent-level β-susceptibility across value dimensions under different openness persona prompts (Qwen3-8B). Each bar corresponds to one value and three colors show β under high-, neutral-, and low-openness personas. While high-openness generally increases susceptibility, the magnitude of this effect varies across values.

orientation in response to peer signals under fixed interaction conditions. We focus on agent-level behavior to isolate intrinsic responsiveness independent of network structure.

Concretely, we consider controlled settings in which a target agent observes a fixed number of preceding responses whose value orientations are systematically perturbed. Let $\bar{x}_i$ denote the average value orientation score of the agent's input context under perturbation configuration i, and let y_i denote the agent's resulting output value score. We empirically observe an approximately linear relationship between y_i and $\bar{x}_i$ across perturbation configurations. Therefore, we define *β-susceptibility* by fitting a linear model

$$y_i = \beta \bar{x}_i + c + \epsilon_i, \tag{3}$$

and interpreting the slope β as the agent's intrinsic sensitivity to peer value signals. A larger β indicates that unit changes in aggregated input values induce larger shifts in the agent's output, reflecting higher susceptibility to value perturbation. Details of β-susceptibility are provided in Appendix E.

2.4.2 System-Level Susceptibility

While agent-level susceptibility captures local response behavior, system-level susceptibility measures how value perturbations propagate through an interacting system under a fixed agent configuration. Here, we vary network topology and perturbation location while fixing agent behavior.

Let y_v^{base} and y_v^{pert} be the value scores of output node v without and with perturbation respectively. We define *system susceptibility (SS)* as

$$SS = \frac{1}{|O|} \sum_{v \in O} \frac{|y_v^{\text{pert}} - y_v^{\text{base}}|}{\Delta_{\text{pert}}}, \tag{4}$$

where O is the set of output nodes and Δ_{pert} is the magnitude of the injected perturbation.

SS quantifies the average impact of a localized unit value perturbation on final system outputs. By normalizing with respect to perturbation strength, SS enables direct comparison across different network topologies and perturbation locations.

3 Measuring Agent-Level Value Perturbation with Susceptibility

We first study agent-level susceptibility in isolation to characterize how a single agent responds to input value perturbations under controlled interaction conditions, independent of network structure. Specifically, we analyze how agent-level β-susceptibility varies across value dimensions and experimental factors, providing a foundation for the system-level analyses in Section 4.

3.1 Experimental Setup

We measure agent-level susceptibility in a controlled setting that isolates agent response behavior from network structure. A single target agent observes responses from a fixed set of preceding agents ($n = 5$), among which a subset is perturbed toward one extreme of the evaluated value dimension, using the strategy in Section 2.3. The target agent's response is scored using the value quantification procedure in Section 2.2, and susceptibility is measured using β-susceptibility (Section 2.4.1).

Within this setup, we vary four factors: **(i) evaluated value dimension** (56 SVS values), **(ii) backbone model** (Qwen3-8B, LLama-3.3-70B, GPT-3.5-Turbo, GPT-4o, Gemma-3-27B; $t = 0$), **(iii) openness persona** (high, neutral, low; detailed prompts shown in Appendix D.3), and **(iv) input**

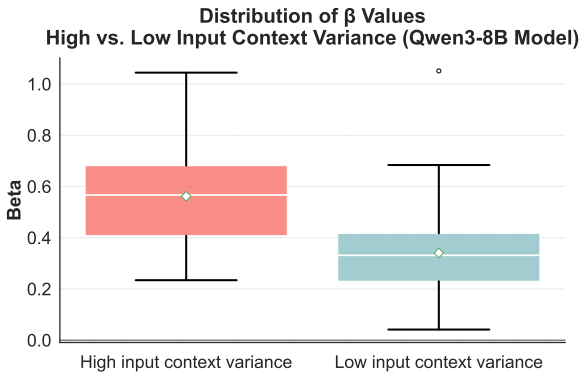

Figure 5: Distribution of agent-level β-susceptibility under high and low input context variance (Qwen3-8B, neutral openness persona). Each box summarizes β values across all 56 value dimensions.

variance (low vs. high). Unless otherwise specified, Qwen3-8B, neutral openness, and high input variance are used as defaults. All other variables are held constant.

3.2 Variation Across Value Dimensions

We first examine how agent-level susceptibility varies across value dimensions under a fixed agent configuration (Qwen3-8B, neutral openness). For each of the 56 values defined in the Schwartz Value Survey, we compute the β-susceptibility of the target agent. Figure 3 shows substantial variation in β across values, indicating that susceptibility is highly non-uniform.

Observation Values that are broadly normative or widely shared (e.g., *Social Power*, *True Friendship*, *Self-Discipline*) consistently exhibit low β, while more context-dependent or socially contingent values (e.g., *Preserving Public Image*, *Influential*, *Detachment*) show substantially higher β.

Finding 1: Agent-level susceptibility varies substantially across value dimensions. Generally, widely shared normative values exhibit low susceptibility, while context-dependent values show significantly higher susceptibility under identical agent configurations.

3.3 Effect of Openness Persona

We next examine how openness persona prompting modulates agent-level susceptibility. For each value dimension, we compute β under low-, neutral-, and high-openness prompting.

Observation Figure 4 shows that β generally increases with openness, indicating amplified sensitivity to peer value signals. However, the magnitude of this increase varies substantially across values. Some values remain stable across persona configurations, while others exhibit sharp contrasts in β under different openness prompting, even when baseline susceptibility is low.

Finding 2: Openness persona prompting **selectively** amplifies agent-level susceptibility for specific value dimensions, rather than uniformly increasing responsiveness.

3.4 Effect of Backbone Models

We evaluate agent-level susceptibility across five backbone models under neutral openness prompting. We document the average β-susceptibility across all 56 values for each backbone, as shown in Table 1. Detailed value-wise results are provided in Appendix F.

Observation Across all 56 value dimensions, backbone models differ substantially in the overall scale of β-susceptibility, with Gemma3-27B and Qwen3-8B exhibiting higher average β, whilst GPT-4o and LLama-3.3-70B are showing lower β-susceptibility.

Table 1: Average agent-level β-susceptibility across backbone models. All values are computed under neutral openness persona over 56 value dimensions.

Backbone Model	Mean β
Gemma3-27B	0.6050
Qwen3-8B	0.5620
GPT-3.5-Turbo	0.4515
GPT-4o	0.4078
LLama-3.3-70B	0.3245

Finding 3: Agent-level susceptibility varies systematically across backbone models.

3.5 Effect of Input Variance

We then examine how variance in inputs affects β-susceptibility, holding all other factors constant.

We compare two settings. In the **low-variance setting**, all preceding agents are instantiated with identical contextual prompts, resulting in highly

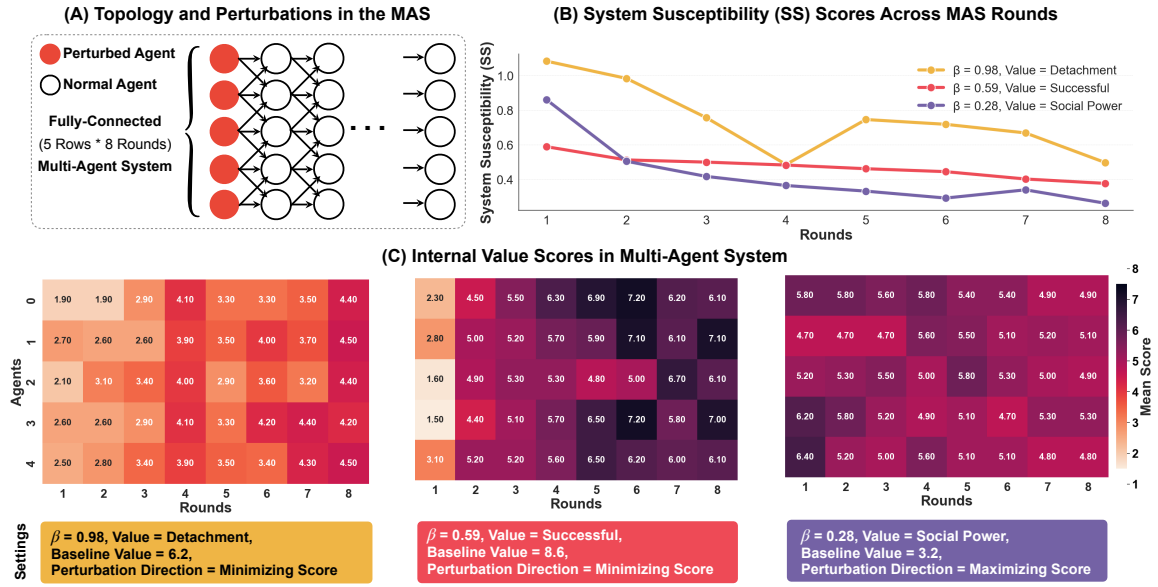

Figure 6: System-level value propagation under different agent-level susceptibility regimes in a fully-connected topology. **(A)** Experimental setting: a fully-connected multi-agent system where all agents in the first round receives value perturbations, and remaining layers follow the original protocol. **(B)** System susceptibility (SS) as a function of interaction depth for three representative values with high, medium, and low agent-level β-susceptibility. **(C)** Evolution of internal value scores across agents and rounds in each three cases of β. Higher agent-level susceptibility leads to slower attenuation of perturbations and more persistent system-level deviations, whereas low-β values exhibit rapid decay toward baseline.

similar value-oriented responses; whereas in the **high-variance setting**, preceding agents are instantiated with distinct contextual prompts, which produce value-consistent but diverse responses. Importantly, the average input value signal $\bar{x}$ is mostly matched across the two settings. The distribution of the βs are shown in Figure 5. Detailed implementation of contextual prompts are in Appendix D.5.

Observation High input variance produces a consistent upward shift in the β distribution across value dimensions, reflected in both the median and interquartile range.

Finding 4: Input variance significantly increases agent-level susceptibility, even when average input value signals are held constant.

4 Measuring System-Level Value Propagation in Multi-Agent Systems

In this section, we examine how value perturbations propagate at the system level in multi-agent networks. Building on the agent-level susceptibility analysis in Section 3, we study how intrinsic agent responsiveness and network structure jointly shape system-level outcomes. Specifically, we address two questions: (i) **how agent-level susceptibility translates into system-level propagation under fixed topology**, and (ii) **how topology and perturbation location affect the magnitude and persistence of value propagation**.

4.1 Effects of Agent-Level Susceptibility on System-Level Propagation

To isolate the effect of agent-level susceptibility, we construct a layered, fully-connected multi-agent system with a fixed number of agents ($n = 5$) per layer. Value perturbations are injected only at the first layer, while subsequent layers follow the standard interaction protocol. All agents are assigned distinct contexts. The backbone model, persona configuration, and topology are held constant. We track internal value scores across interaction rounds and measure system susceptibility (SS) as a function of network depth. We compare three representative values with high, medium, and low agent-level susceptibility ($\beta = 0.98, 0.59, 0.28$).

Observation Figure 6 shows that agent-level susceptibility strongly predicts system-level propagation. Higher-β values exhibit persistent deviations across layers and slow decay in SS, whereas lower-β values attenuate rapidly and converge toward