He Said, She Said, They Simulated: Simulating Situated Gossip with LLM Agents

Yicong Tan[1], Yun Shen[2], Yang Zhang[1]*
[1]CISPA Helmholtz Center for Information Security, [2]HPE
yicong.tan@cispa.de, yun.shen@hpe.com, zhang@cispa.de

Abstract

Gossip has been investigated primarily through vignette-based surveys, which provide limited insights into the dynamic and adaptive gossip behaviors within real-world contexts. To address these limitations, we introduce an LLM agent-based simulation framework, designed as a configurable experimental laboratory for gossip research. We validate our framework across three core dimensions of gossip dynamics: initiation, reaction, and perception. Our framework can reproduce several behavioral trends observed in human studies and enable analysis of agents' reasoning and temporally unfolding interactions. We further augment the framework with a chain-of-reasoning mechanism to support exploratory investigation, demonstrating how explicit agent reasoning and experiment-level controllability enable the generation of hypotheses about gossip dynamics beyond fixed survey settings. Overall, our framework bridges LLM-based computational modeling and social science, complementing vignette-based methodologies and enabling controlled, systematic study of complex social behaviors[1].

1 Introduction

Gossip, an informal communication about absent others, is a ubiquitous but controversial element of human social life (Michelson et al., 2010). It plays an important role in regulating cooperation and competition (Wu et al., 2016; Wyckoff et al., 2019), reforming social norms (Giardini and Conte, 2012), and boosting work performance (Bai et al., 2020). At the same time, gossip also carries risks; it can spread negative information (McAndrew et al., 2007), promote social exclusion (Martinescu et al., 2021), and lead to psychological harm (Wu et al., 2018). This dual nature has made gossip a key lens through which researchers investigate complex social cognition and group dynamics (Kurland and Pelled, 2000; Beersma and Kleef, 2012; Brady et al., 2017; Cruz et al., 2021).

Figure 1: Gossip simulation in the sandbox environment. We focus on a small-scale environment that resembles real-world office layouts.

Despite their importance, most gossip studies are based on vignette-based surveys where participants respond to hypothetical scenarios (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al., 2021). While these methods offer controlled yet static environments to probe people's responses, they fail to capture the context-sensitive aspects that characterize real-world gossip interactions. As a result, we lack tools to study how gossip emerges, adapts, and propagates within dynamic real-world environments, limiting both theoretical development and exploratory insights.

To address this gap, we propose an agent-based simulation framework (see Figure 1) in which Large Language Model (LLM) agents engage in routine social interactions within a simulated workplace environment. Equipped with agent reasoning, memory, and language abilities demonstrated in prior work (Park et al., 2023; Xie et al., 2024; Piao et al., 2025; Wu et al., 2025), these agents

*Corresponding Author.
[1]Our code is available at: https://github.com/TrustAIRLab/Gossip-Simulacra

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 36763–36806
October 24-29, 2026 ©2026 Association for Computational Linguistics

enable us to instantiate, observe, and test the dynamics of gossip as it naturally occurs from situated interactions. Unlike existing large-scale LLM agent simulations (Hu et al., 2025; Piao et al., 2025; Tang et al., 2024; Wang et al., 2025) that model population-level information diffusion or social media dynamics, *our approach focuses on capturing the fine-grained and situated gossip behaviors that emerge between individuals during routine office interactions.*

To validate our framework, we compare the simulation results with established findings in social psychology (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al., 2021). The goal is to assess the fidelity of LLM-based simulations in approximating known behavioral patterns. Our results in Section 4.3 show that LLM agents can exhibit similar behavioral patterns and, at the same time, generate additional hypotheses by enabling the modeling of dynamic, real-time gossip behaviors that reflect everyday social interactions within the workspace.

Beyond validation, we introduce customizable gossip experiments that extend the framework toward open-ended exploratory research rather than approximation. *We augment the framework with a structured chain-of-reasoning that decomposes agent behavior into sequential decision stages*, guiding agents toward gossip-related interaction while preserving autonomy over when to gossip, whom to target, and what content to share. This balance between experimental control and agent autonomy enables the generation of hypotheses about gossip dynamics beyond existing studies, as shown in Section 5.

Our contributions are summarized below:

- To the best of our knowledge, we are the first to model situated gossip dynamics using LLM agent-based simulation.
- We empirically assess the behavioral alignment of LLM agents with existing findings from gossip studies.
- Extensive experimental results demonstrate that LLM agents supporting customizable gossip simulations can generate hypotheses about emergent gossip behaviors.

2 Related Work

Vignette. Vignette-based surveys are a common methodological approach in gossip research, wherein participants are asked to imagine how they would respond to hypothetical social scenarios (McAndrew et al., 2007; Beersma and Kleef, 2012; Martinescu et al., 2019; Tan et al., 2021). While such methods offer controlled insights into social reasoning, they often rely on static, oversimplified scripts and do not capture the real-time dynamics of interpersonal interaction, which are deeply intertwined with everyday social interactions within groups (Dunbar, 2004; Foster, 2004). For instance, (Martinescu et al., 2019) presents participants with an imagined group work scenario involving overheard gossip and then measures emotional and behavioral responses.

LLM Agents. The powerful capabilities of LLM-based agents have expanded the methodological toolkit of social science research (Park et al., 2023; Tang et al., 2024; Piao et al., 2025), enabling realistic simulations of human behavior. LLMs have demonstrated the capacity to approximate human-like cognition, such as theory of mind (Strachan et al., 2024), general reasoning (Bubeck et al., 2023), moral belief (Scherrer et al., 2023), emotional intelligence (Schlegel et al., 2025), cooperation (Zhang et al., 2024), and emergent social behaviors and conventions (Leng and Yuan, 2023; Ashery et al., 2025), thereby underscoring their potential for investigating complex social phenomena. This facilitates frameworks that support social simulation with LLM agents, including generative agents (Park et al., 2023), YuLan-OneSim (Wang et al., 2025), GenSim (Tang et al., 2024), and Agent Society (Piao et al., 2025). At the same time, LLM agents have supported domain-specific social simulations ranging from behavioral studies in game-theoretic settings (Lan et al., 2024; Abdelnabi et al., 2024; Mao et al., 2025; Buscemi et al., 2025) to modeling human-like behavior in real-world-inspired settings (Park et al., 2023; Wu et al., 2024; Xie et al., 2024; Hou et al., 2025).

3 Simulation Framework

3.1 Overview

Existing methods (Park et al., 2023; Tang et al., 2024; Piao et al., 2025) primarily model social interaction as an emergent property of agent behavior. They offer limited control over the environment to elicit specific gossip interactions, which may hinder the study of how gossip emerges, adapts, and propagates within dynamic real-world environments. *To address this gap, we propose a framework that*

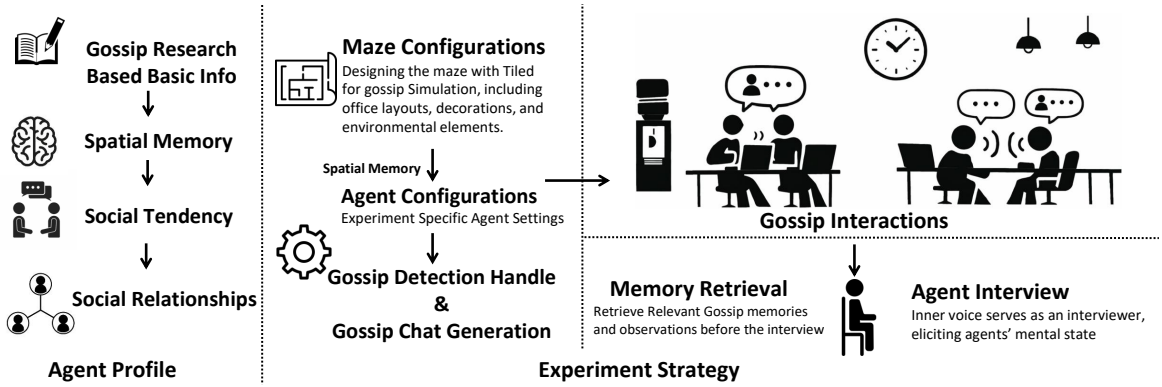


Figure 2: Gossip simulation framework with socially grounded agent profiling and modular experimental strategy.

incorporates socially grounded agent profiles, predefined interpersonal relationships, and a modular experimental strategy formulated specifically to simulate, control, and evaluate gossip-related interactions. These modules, as shown in Figure 2, enable controlled, context-aware gossip simulations, supporting the study of gossip dynamics with greater sociological fidelity and generating additional insights in real-world settings. The differences from generative agents (Park et al., 2023), which our framework builds on, are discussed in Appendix H.1.

Socially Grounded Agent Profile. The agent profile shapes each agent's social and behavioral context and consists of **four** components. *Gossip research-based basic information* captures attributes such as age, identity, occupation, and scratch information, including background descriptions and lifestyle details, closely matching the original study settings by grounding profiles in participant data from prior work. *Spatial memory* represents agents' knowledge of the office layout and the locations of other agents. *Social tendencies* include social habits and conversational intent, which regulate when and how agents engage in social interactions. *Social relationships* specify agents' relational ties to others, reflecting the relationship categories defined in the original studies.

Modular Experimental Strategy. We introduce a configurable experimental strategy to guide agent interactions in the gossip simulation, enabling controlled setup of agent characteristics and systematic analysis of agents' responses to gossip events. The experimental strategy consists of **four** components. *Agent configuration* initializes experiment-specific agent attributes. *Gossip detection handle* manages agent reactions to gossip events. *Gossip chat generation* produces dialogue conditioned on gossip simulation context. *Agent interview* extracts post-simulation agent behavioral outcomes by providing the agent with memories accumulated through gossip interactions. Intermediate interactions are also recorded to support analysis of how gossip propagates across agents. These components support controlled investigation of gossip initiation, response, and information spread within a configurable agent environment.

4 Validation against Prior Work

We first apply this framework to three established areas of gossip research, namely gossip initiation, gossip reaction, and gossip perception, to assess whether our framework can reproduce established empirical findings from gossip research.

4.1 Gossip Studies

As illustrated in Figure 3, gossip, as a social activity of talk, fundamentally involves two complementary aspects: how individuals speak about others and how they hear, interpret, and respond to others' talk.

Study 1: Gossip Initiation. Gossip initiation (McAndrew et al., 2007) examines: *How do social proximity and the initiator-recipient relationship influence recipient selection across different types of gossip?* and *How do gossip valence (positive vs. negative) and the gossip initiator-main character relationship jointly shape recipient choice?*

Study 2: Gossip Reaction. Gossip reaction (Martinescu et al., 2019) investigates: *How does gossip valence influence self-conscious and other-directed emotions, as well as behavioral intentions, in various scenarios?*

Study 3: Gossip Perception. Gossip perception (Tan et al., 2021) explores: *How does expo-*

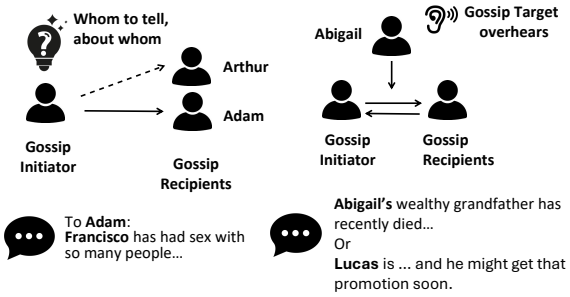


Figure 3: Gossip is either intentionally initiated toward selected recipients or overheard by nearby agents, with gossip about others or oneself.

sure to high (vs. low/control) levels of work-related gossip affect perceived performance pressure and psychological well-being?

4.2 Experiment Settings

Experimental Setup. We closely adhere to the settings of prior gossip studies (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al., 2021) to construct agent profiles and gossip content. Details of agent profile construction and prompt templates for the experiment strategy are provided in Appendix A and Appendix B. Gossip reaction and perception are conducted with 64 agents, and gossip initiation with 16 agents per run, simulated over a single workday from 9:00 a.m. to 5:00 p.m. using 10-minute time steps. The framework is evaluated with both open-source and closed-source LLM backbones, including GPT-4.1-mini (OpenAI, 2025), Llama-3.1-8B, Llama-3.3-70B (Team, 2024), and Gemma-3-27B (Team, 2025), deployed via the Ollama (Ollama, 2026) or vLLM framework (Kwon et al., 2023). Detailed environment construction, model configurations, experiment repetition, and runtime parameters are provided in Appendix C.

Evaluation Metrics. For comparability, we use the original study metrics: ranking and probability measures for gossip initiation, and 7-point Likert scales for reaction and perception. Equivalence is pre-specified using two one-sided tests (TOST) (Schuirmann, 1987), with a margin of one human standard deviation to reflect human response variability (Xi and Jackson, 2025) and variation in prior gossip studies detailed in Appendix C.3.

4.3 Results

We conduct LLM agent-based simulations to approximate the gossip behaviors reported in prior studies (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al., 2021). Additional analyses that probe agent reasoning when behaviors diverge from established patterns are reported in Appendix D.

4.3.1 Gossip Initiation

As shown in Figure 7 in Appendix D.1, LLM agents in our framework successfully simulate gossip recipient selection by preferentially sharing gossip with allies such as friends, relatives, and acquaintances (rank top-3) rather than with strangers and rivals (rank 8-12), closely matching reported conclusions in existing research. In addition, as shown in Table 12 in Appendix D.1, LLM agents also partially approximate gossip content selection patterns: they consistently favor sharing positive information about allies (probability ≈ 1), while preferences for sharing negative information about rivals vary across LLMs, showing model-dependent differences. Additional analyses, model comparisons, and reasoning for explaining the observed model-dependent differences are reported in Appendix D.1.

4.3.2 Gossip Reaction

LLM agents approximately simulate gossip reaction patterns across two scenarios from (Martinescu et al., 2019) when exposed to positive or negative gossip. As shown in Figure 8 in Appendix D.2, simulated agents exhibit emotional responses closely aligned with human evaluations, particularly for self-conscious emotions (both for negative $\approx$ 3.0 and positive $\approx$ 4.0) and other-directed emotions (both for negative $\approx$ 3.0-3.5 and positive $\approx$ 3.0-4.0). Simulated behavioral responses also broadly match human patterns, though agents tend to show weaker negative ($\approx$ 1.5 vs. human $\approx$ 2.0) and stronger positive behavioral intentions in repair ($\approx$ 5.5-6.0 vs. human $\approx$ 4.5-5.0), with variable affiliation patterns ($\approx$ 1.5-5.0 vs. human $\approx$ 3.8). Importantly, this alignment is not merely numerical: as shown in Table 14 and Table 16, TOST (Schuirmann, 1987) shows equivalence to humans within one standard deviation (average minimal margin around 1.10 compared to human standard deviation 1.77). Additional analyses, statistical tests, and elicited reasoning are reported in Appendix D.2.

4.3.3 Gossip Perception

LLM agents approximately simulate the psychological impacts of perceived work-related gossip prevalence across control and high-gossip conditions from (Tan et al., 2021). As shown in Table 17 in Appendix D.3, simulated agents produce ratings closely aligned with human evaluations for perceived gossip prevalence (control ≈ 4.0 vs. human 4.3; experiment ≈ 5.0 vs. human 4.9) and performance pressure (control ≈ 5.7 vs. human 5.5; experiment ≈ 5.8 vs. human 5.8). For psychological well-being, agents and humans both report well-being ratings (the higher the worse) decreasing from the high-gossip condition experiment to control (agents $\approx 3.9 \rightarrow 3.5$ versus humans $4.85 \rightarrow 4.34$). Moreover, TOST (Schuirmann, 1987) confirms the equivalence to humans within one standard deviation (minimal margin around 0.7 compared to human standard deviation 1.2). Additional analyses of variability, model-specific effects, statistical tests, and elicited reasoning are reported in Appendix D.3.

4.3.4 Gossip Simulation Robustness

Beyond replicating, we introduce factors absent from the original human studies: agent profiles, workspace roles, memory valence, and environment, and show that they yield diverse, context-sensitive behavioral shifts in Appendix E, illustrating the framework's potential for open-ended exploration. Our simulations also exhibit dynamic gossip behaviors similar to daily life that emerge through ongoing conversations and context-dependent reasoning, including varied framing, follow-up interactions, and multi-step propagation across agents. Detailed conversational examples and agent reasoning traces are provided in Appendix D.6. We also show an IRB-approved human comparison study in Appendix D.5.

5 Exploratory Capability Assessment

Having established that LLM agents reproduce several empirically observed gossip behaviors, we next use the framework as a controlled experimental platform to explore *how organizational structure and personality traits shape gossip dynamics in the context of peer review.* This experiment highlights the framework's customizability through *agent-level reasoning* and *experiment-level controllability* to generate hypotheses about gossip behavior.

5.1 Overview

Background. Agreeableness is one of the five fundamental dimensions of personality (Costa and McCrae, 2008) and captures tendencies toward cooperation, altruism, and interpersonal harmony. Peer review is a common evaluative process in which colleagues assess one another's work performance. When combined with agreeableness, this evaluative context offers a setting for exploring workplace gossip behavior.

Setup. To investigate this, LLM agents with varying levels of agreeableness are placed into a shared workspace. Agents are informed that it is the final day before the peer-review deadline, made aware of the current sales performance rankings, and told that the final ranking will influence their bonuses. We observe agents' interactions throughout this final day. At the end of the simulation, agents retrieve stored interactions and observations from memory to generate peer-review evaluations.

Agent Profile. We adapt the agent profile settings in Section 3.1. To construct a controlled experimental environment, we vary agent profiles only along our research variables, such as agreeableness, workspace roles, and team membership, while keeping all other attributes constant across agents. This ensures that any behavioral differences can be attributed to these targeted factors. Details of agent profiles are provided in Appendix F.1.

Experiment Strategy. We follow the design scheme outlined in Section 3.1 to construct the experiment strategy for this more customizable setting. However, motivated by the lack of interactions and subjective initiative shown in our validation simulations, we introduce a structured chain-of-reasoning mechanism that guides agents through each stage of the gossip process. This stepwise elicitation improves the coherence and intentionality of simulated behaviors, as illustrated in Figure 4. Furthermore, we conduct ablation experiments of this mechanism in Appendix H.3. Consequently, agents in our simulations are guided to make coherent, sequential decisions rather than produce isolated, event-specific reactions.

5.2 Gossip Studies

Unlike the validation experiments, which verify psychological principles against human studies, the peer-review setting is designed not for validation but for hypothesis generation: agents engage in gossip under competitive incentives, pro-

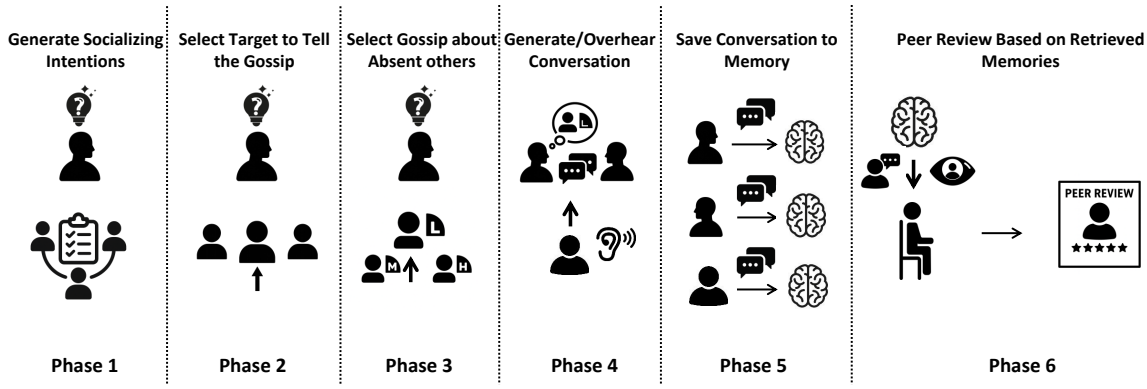


Figure 4: Chain-of-reasoning for customizable gossip simulation, extending the validation setting by replacing fixed socializing intent with situationally grounded, sequential gossip decision-making with details in Appendix F.2.

ducing context-dependent patterns that extend beyond fixed vignettes. To study this, we design four customizable experiments that examine how agreeableness shapes gossip behavior under peer-review incentives. These experiments introduce controlled variations in social and organizational conditions while maintaining a common simulation backbone. The customizable modifications are summarized in Table 30 in Appendix F.2. We first consider a (1) **baseline experiment** where all agents perform equivalent roles and compete individually for performance-based bonuses determined by peer-review scores. The study aims to understand: *How do agents strategically select gossip targets and valence in a flat organizational structure where competition is purely individual?* We then introduce additional social and organizational factors to examine their influence on gossip behavior. These *extensible gossip experiments* include: (2) **hierarchical roles**, where agents have unequal influence in peer review; (3) **team-based group structures**, enabling intra- and inter-group evaluation; and (4) **two-wave interactions**, allowing gossip effects to unfold over time. The hierarchical setting examines how asymmetric influence affects gossip strategies. The team-based structure compares behaviors within and across groups, revealing how group boundaries shape gossip behavior. In the two-wave interaction setup, simulations occur in two phases separated by a one-month interval. Agents receive summarized gossip from the first phase before the second begins, allowing reputation signals to influence subsequent behavior and evaluations.

5.3 Experiment Settings

We follow the simulation environment settings in Section 4.2, with targeted modifications for the customizable gossip experiments. We construct five experimental conditions that vary agents' agreeableness (scores 1-5): homogeneous low (**Low-All**, scores $\leq$ 2.5), homogeneous high (**High-All**, scores $>$ 3.5), **Balanced** (32 low and 32 high), **Gradient** (agents evenly distributed across scores 1-5), and **Normal** (scores drawn from a normal distribution). These settings allow us to examine how agreeableness influences workplace gossip patterns. Agreeableness scores are translated into textual descriptions to provide agents with contextual cues, as shown in Table 31 in Appendix F.3. Detailed configurations for the base and extended gossip experiments are provided in Appendix F.3. We use GPT-4.1-mini for these simulations due to its stability and low cost. Since our goal is to evaluate framework extensibility rather than model variability, a single lightweight model is sufficient.

5.4 Results

Rather than offering an exhaustive analysis of agent behaviors, we focus on experiment results to demonstrate how *agent-level reasoning* and *experiment-level controllability* jointly enable our framework to support *hypothesis generation* about gossip dynamics.

5.4.1 Base Gossip Experiment

Agent Reasoning. Agent chain-of-reasoning refers to a structured, agent-level decision process that decomposes social behavior into sequential steps and should not be confused with the underlying language model's reasoning capability. As shown in Figure 4, agent chain-of-reasoning in the baseline gossip setting is illustrated across multiple phases. For instance, in Table 32 in Appendix G, an agent with moderate-to-low agreeableness (Ayesha, 2.4) reasons through a multi-stage decision pipeline

during the baseline gossip experiment. This illustrates how agent-level reasoning unfolds sequentially, beginning with the generation of a daily socializing intent that integrates personality traits, competitive incentives, and peer-review rules. This intent then conditions subsequent interaction decisions, including whether to initiate a conversation, which nearby agent to approach, how to ground the interaction through relationship summarization, and which absent colleague and evaluative angle to select for gossip. Finally, the agent generates gossip content that is explicitly aligned with the chosen angle and contextual constraints. Consequently, this process illustrates a coherent agent-level reasoning pipeline that maps internal goals to observable gossip behaviors.

Experiment Controllability. Gossip is inherently the art of language use, shaped by subtle linguistic choices rather than predefined actions. Accordingly, our baseline gossip experiments enforce controllability through a structured, prompt-based simulation lifecycle spanning profiling, interaction, and post-evaluation, while leaving agents' autonomy to realize gossip linguistically. Across the simulation, prompts and structured interactions impose consistent experimental constraints and provide relevant context that focuses agents on gossip-related interaction, while preserving autonomy over how gossip is expressed through language. This design elicits gossip behavior without prescribing specific utterances or strategies. The controllability is empirically reflected in both agents' linguistic behavior and peer-review results. First, gossip valence shows a strong dependence on agreeableness. Using keyword-based categorization of gossip content, Table 1 and Table 36 in Appendix G show that low-agreeableness groups produce negative gossip (79.0%), whereas high-agreeableness groups generate more positive gossip (98.2%). Second, controllability in gossip behavior allows us to jointly examine the effects of gossip initiation and being gossiped about within a single experimental framework. As shown in Table 2, this control produces context-dependent patterns in peer-review outcomes, enabling direct comparison of how initiating gossip and receiving gossip relate to rank change across agreeableness conditions. **In addition**, we evaluate *a conflict personality setting in which an agent combines high agreeableness and high neuroticism to assess the agent's ability to adapt to competing personalities.* Across repeated simulations, the agent exhibits coherent adaptive behavior, maintaining predominantly positive interactions while intermittently expressing negative gossip, as reflected in both socializing intent generation and conversation valence, as shown in Table 33 and Table 37 in Appendix G. This indicates that agent behavior emerges from personality traits interacting with situated social context, not fixed prompts or profiles (Appendix H.4 further isolates the personality wording and prompt effects).

Condition	Positive	Negative	Total
Low-All	66.4 (21.0±1.6%)	249.4 (79.0±1.6%)	315.8
High-All	390.2 (98.2±1.1%)	7.2 (1.8±1.1%)	397.4
Balanced	208.2 (58.6±3.2%)	147.0 (41.4±3.2%)	355.2
Gradient	234.8 (65.4±2.4%)	124.4 (34.6±2.4%)	359.2
Normal	353.0 (87.5±1.4%)	50.2 (12.5±1.4%)	403.2

Table 1: Gossip valence results across agreeableness conditions with agents equipped with reasoning, further classification by LLM-as-judge in Appendix G.3.

Condition	Gossip Initiation		Being Gossiped About	
	Positive	Negative	Positive	Negative
Low-All	-0.073	$+0.019$	$+0.112$	-0.162^*
High-All	$+0.021$	$+0.039$	$+0.256^*$	$+0.131^*$
Balanced	-0.008	-0.047	$+0.210^*$	-0.109
Gradient	-0.050	$+0.036$	$+0.298^*$	-0.032
Normal	-0.125^*	$+0.001$	$+0.311^*$	$+0.039$

Table 2: Repeated-measures correlation (Bakdash and Marusich, 2017) between gossip and peer-review rank improvement across agreeableness conditions. * indicates statistical significance ($p < 0.05$).

Generated Hypotheses. We further report hypotheses about previously unexplored gossip dynamics enabled by agent reasoning. First, *experimental control preserves linguistic autonomy, allowing the framework to generate gossip that is open to interpretation*, as shown in Table 34 in Appendix G.1, pointing toward future work on analyzing nuanced gossip phenomena. Second, *we identify a limited social signaling effect*: although gossip initiation is often regarded as socially controversial in prior work (Michelson et al., 2010), our results suggest a weaker effect. As shown in Table 2, the social signaling role of gossip initiation may emerge only under specific personality compositions rather than as a general property. In contrast, being gossiped about, particularly positively, is associated with rank improvement in four of the five conditions. These hypotheses demonstrate the customizability of the framework to uncover various gossip dynamics.

Agreeableness Range / Hierarchical Role	Junior				Mid-level				Senior			
	Up %	Down %	Same %	number	Up %	Down %	Same %	number	Up %	Down %	Same %	number
Very Low ($\leq$ 2.0)	78.8±2.0	0.0±0.0	21.2±2.0	239.3	22.9±3.0	16.8±1.7	60.3±2.0	262.3	0.0±0.0	75.3±8.3	24.7±8.3	142.3
Low (2.0–2.5)	83.1±4.5	0.0±0.0	16.9±4.5	38.0	17.9±15.8	15.7±4.0	66.4±12.2	53.7	0.0±0.0	69.8±5.8	30.2±5.8	83.3
Medium (2.5–3.5)	87.4±3.9	0.0±0.0	12.6±3.9	180.7	25.6±5.5	16.9±1.5	57.5±5.5	148.7	0.0±0.0	66.9±4.8	33.1±4.8	75.0
High (3.5–4.5)	70.7±2.0	0.0±0.0	29.3±2.0	136.7	11.3±4.6	15.0±3.0	73.7±1.9	212.7	0.0±0.0	78.5±6.1	21.5±6.1	188.0
Very High ($>$ 4.5)	79.7±1.0	0.0±0.0	20.3±1.0	87.0	20.9±5.6	16.7±7.4	62.5±10.8	101.3	0.0±0.0	72.1±10.2	27.9±10.2	75.3
Overall Negative Percentage	**Upward: 35.6±1.4%**				**Downward: 37.3±5.5%**				**Same-Level: 34.6±1.5%**			

Table 3: Distribution of gossip direction across agreeableness ranges and hierarchical roles. Percentages indicate the proportion of gossip directed upward, downward, or toward same-level colleagues within each role; $number$ denotes the total number of gossip instances. We also show the overall negative gossip percentage for agents of all hierarchical roles.

5.4.2 Extensible Gossip Experiments

Agent Reasoning. Agent reasoning in the extensible gossip settings demonstrates context-sensitive adaptation to different extensions. As illustrated in Table 39 in Appendix G.2, in the hierarchical role setting, agents explicitly reference organizational rank when evaluating colleagues, reasoning about senior complacency, and positioning themselves for upward mobility. In the team-based group setting, agents incorporate group identity and inter-group competition, coordinating gossip strategies to suppress rival teams' peer-review outcomes while protecting in-group members. In the two-wave setting, agents exhibit temporal reasoning by adjusting their socializing intent after learning the first-wave gossip, strategically targeting those agents in subsequent interactions. These findings indicate that the reasoning pipeline flexibly integrates *hierarchical awareness*, *group structure*, and *first-wave gossip* to produce contextually relevant gossip behaviors.

Experiment Controllability. The extensible gossip experiments illustrate *how experiment controllability extends beyond the base setting, as the framework supports the controlled integration of hierarchy, group structure, and two-wave simulation without altering the underlying reasoning pipeline*. As shown in Table 3, the framework enables controlled variation of agreeableness and hierarchical roles, yielding distinct distributions of upward, downward, and peer-level gossip across hierarchical positions. Table 4 demonstrates that controlled manipulation of team structure and agreeableness distributions produces both in-team and cross-team gossip. As shown in Table 38 and Table 40 in Appendix G.2, agents' understanding of gossip information in the first wave leads to increased gossip activity in the second wave (an 81% to 134% volume increase across the five agreeableness conditions), with subsequent peer-review ratings change manifesting as both gratitude and retaliation depending on agreeableness (Low-All retaliate at 27% of pairs versus 5% for High-All).

(a) In-team gossip rate by initiator agreeableness

Agreeableness	In-team %	Cross-team %	Number
Very Low	23.73±4.20	76.27±4.20	806
Low	22.00±3.01	78.00±3.01	339
Medium	25.30±4.67	74.70±4.67	387
High	25.84±3.13	74.16±3.13	621
Very High	25.04±2.18	74.96±2.18	292
Random-target baseline: 23.8%			

(b) Sentiment of in-team vs. cross-team gossip

Direction	Negative %	Positive %	Number
In-team	26.01±1.83	70.23±1.44	596
Cross-team	**28.29±2.93**	68.50±3.09	1849
Difference	**+2.27±4.74**	**-1.73±4.32**	—

Table 4: Team-based gossip patterns.

Generated Hypotheses. We report hypotheses beyond the controlled experimental results. In the hierarchical role, *agents exhibit gossip behavior that is not deterred by authority*, continuing to direct upward gossip (e.g., 80% for very high agreeableness agents), as shown in Table 3. For the team-based extension, as shown in Table 4, we observe that team identity can be expressed through gossip semantics rather than target choice: cross-team gossip is slightly more negative than in-team gossip (28% versus 26% negative), while in-team gossip rates sit near the random-target baseline of 23.8% across all agreeableness levels. For the two-wave extension, agents under Low-All conditions gossip back at prior gossipers with overwhelmingly negative gossip in Wave 2 (95.6% negative), while agents under High-All conditions respond with overwhelmingly positive gossip-back (only 5.2% negative), indicating that sentiment dynamics across waves are governed by personality composition, as shown in Table 5.

Framework Extensibility. Our framework also supports modular extensions to the agent architecture (psychological state, inner monologue, visible

Condition	Wave-2 Gossip-Back Instances				Neg%
	Negative	Positive	Neutral	Total	
Low-All	735.0	26.0	7.7	768.7	95.6±2.4
High-All	26.3	339.7	142.3	508.3	5.2±2.3
Balanced	249.7	142.0	46.7	438.3	57.0±8.9
Gradient	348.3	165.3	53.3	567.0	61.2±2.6
Normal	64.3	325.3	75.7	465.3	13.8±2.3

Table 5: Valence of Wave-2 gossip-back instances across agreeableness conditions, further classification by LLM-as-judge in Appendix G.3.

facial expression), conversation topology (group chat), and demographic moderators. At the 64-agent scale, visible facial expressions neutralize gossip valence, most strongly for very-low-agreeableness agents, consistent with a social-regulator role, and the group-chat topology yields three-or-more-participant conversations in 69% of all chats. Full details, analysis, and experiment results are in Appendix G.4.

6 Conclusion

In this paper, we simulate gossip dynamics in multi-agent environments. By empirically comparing LLM agent behavior with human evaluations, we assess their alignment in approximating human-like decision-making processes, emotions, behavioral intentions, and psychological reactions. Additionally, we demonstrate the flexibility of our framework in supporting customizable simulations, highlighting the potential of LLM agents for studying complex social interactions. We position our framework as a complement to vignette-based gossip research, enabling the study of emergent multi-agent dynamics that cannot be directly elicited through vignette-based methods. We hope our study will inspire future research on using LLM agents to model complex human social behaviors in real-world settings.

Limitations

Our LLM agent-based simulation framework provides substantial insights into gossip dynamics. However, several limitations inherent to LLMs may constrain their fidelity and generalizability:

Hallucination. In our setting, what may appear as a hallucination often corresponds to agents elaborating on gossip content based on their background and social context, reflecting generative behavior rather than unsupported fabrication. True hallucination cases, such as incorrect assumptions or inconsistent attributions, are rare and typically affect individual interactions rather than overall patterns. Since our analysis focuses on aggregate trends, their impact on the main conclusions is limited.

Bias. The pretraining data of LLMs consists of broad and heterogeneous Internet texts, which inevitably embed various societal biases. As demonstrated in Section 4.3.1, LLM agents guided by different LLMs exhibit divergent reasoning in gossip content selection. These biases can manifest in stereotypical or culturally biased decision-making processes, which may influence the outcomes of the gossip simulations. We elaborate on these discrepancies in Appendix D.4.

Human-LLM Alignment Gap. A residual gap between LLM-agent responses and human respondents remains in our comparisons; full per-condition values appear in Appendix D. We interpret this gap as bounded by, and qualitatively consistent with, the inherent inter-study variability of human gossip research itself. Our validation cross-references three published human gossip studies spanning about 14 years, together with our IRB-approved survey, and observes largely consistent trends across emotional, behavioral, and gossip content and initiation patterns. Exact numerical correspondence is structurally unattainable, in part because the original surveys do not, or cannot, disclose the personalized demographic and contextual metadata required to reconstruct a matched sample. Validation should therefore be read as establishing that LLM agents track human gossip trends within standard equivalence bounds, not as a claim of population-level identity. The customizable factors examined in our exploratory capability assessment in Section 5, including agent profiles, hierarchical roles, team structure, and demographic and cultural backgrounds, are precisely the tools that can be adjusted to strengthen alignment as needed.

Ethical Considerations

IRB. For the human-evaluation comparisons reported in this paper, we additionally collected human responses on the same prompts that were presented to the LLM agents. Participants ($n = 16$, aged 25–28) provided informed consent prior to participation. The collection protocol was reviewed and approved by an Institutional Ethical Review Board (IRB), which concluded that "there are no ethical concerns against the implementation of the proposal." Our LLM-based gossip simulations are conducted solely for research purposes within vir-

tual environments and involve no real-world deployment. All artifacts used in this work comply with their respective licenses.

Misuse. Our framework simulates only fictional personas parameterized from published studies and does not ingest data from identifiable individuals. However, the underlying methodology, i.e., controllable simulation of reputational information flow in workplace settings, is dual-use. If adapted to real employee data, it could support surveillance, targeted reputation manipulation, or optimization of harmful influence strategies. We therefore release code and prompts under a research-only license that explicitly prohibits application to data about real, identifiable persons. The framework is a hypothesis-generating instrument for gossip research, not a predictive tool for real workplaces.

Stereotyping. Agent profiles include demographic attributes, and LLM pretraining biases can cause simulations to reproduce or amplify demographic and cultural stereotypes. Differences in simulated gossip behavior across demographic profiles are rooted in the models, not real demographic groups, and must not be reported as evidence about the latter.

Overgeneralization. Our validation establishes trend-level consistency with three human studies, not population-level equivalence (see Limitations). Simulated findings, including the exploratory results in Section 5, are hypotheses requiring confirmation with human participants before informing any real-world decision.

References

Sahar Abdelnabi, Amr Gomaa, Sarath Sivaprasad, Lea Schönherr, and Mario Fritz. 2024. Cooperation, Competition, and Maliciousness: LLM-Stakeholders Interactive Negotiation. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 83548–83599. NeurIPS.

Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent social conventions and collective bias in LLM populations. *Science Advances*.

Yun Bai, Jie Wang, Tingting Chen, and Fuli Li. 2020. Learning from supervisor negative gossip: The reflective learning process and performance outcome of employee receivers. *Human Relations*.

Jonathan Z. Bakdash and Laura R. Marusich. 2017. Repeated Measures Correlation. *Frontiers in Psychology*.

Bianca Beersma and Gerben A. Van Kleef. 2012. Why People Gossip: An Empirical Analysis of Social Motives, Antecedents, and Consequences. *Journal of Applied Social Psychology*.

Daniel L. Brady, Douglas J. Brown, and Lindie Hanyu Liang. 2017. Moving beyond assumptions of deviance: The reconceptualization and measurement of workplace gossip. *Journal of Applied Psychology*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott M. Lundberg, Harsha Nori, Hamid Palangi, Marco Túlio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. *CoRR abs/2303.12712*.

Alessio Buscemi, Daniele Proverbio, Paolo Bova, Nataliya Balabanova, Adeela Bashir, Theodor Cimpeanu, Henrique Correia da Fonseca, Manh Hong Duong, Elias Fernández Domingos, Antonio Fernandes, Marcus Krellner, Ndidi Bianca Ogbo, Simon T. Powers, Fernando P. Santos, Zia Ush-Shamszaman, Zhao Song, Alessandro Di Stefano, and The Anh Han. 2025. Do LLMs trust AI regulation? Emerging behaviour of game-theoretic LLM agents. *CoRR abs/2504.08640*.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*.

Paul T. Costa and Robert R. McCrae. 2008. The Revised NEO Personality Inventory (NEO-PI-R). *The SAGE Handbook of Personality Theory and Assessment*.

Terence D. Dores Cruz, Annika S. Nieper, Martina Testori, Elena Martinescu, and Bianca Beersma. 2021. An Integrative Definition and Framework to Study Gossip. *Group & Organization Management*.

Robin I. M. Dunbar. 2004. Gossip in Evolutionary Perspective. *Review of General Psychology*.

Robert Eisenberger and Justin Aselage. 2009. Incremental effects of reward on experienced performance pressure: positive outcomes for intrinsic interest and creativity. *Journal of Organizational Behavior*.

Electronic Arts Inc. 2026. The Sims 4. https://www.ea.com/games/the-sims/the-sims-4.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*.

Eric K. Foster. 2004. Research on Gossip: Taxonomy, Methods, and Future Directions. *Review of General Psychology*.

Francesca Giardini and Rosaria Conte. 2012. Gossip for social control in natural and artificial societies. *Simulation*.

Google. 2026. Gemini 3.5 Flash. https://ai.google.dev/gemini-api/docs/models/gemini-3.5-flash.

Abe Bohan Hou, Hongru Du, Yichen Wang, Jingyu Zhang, Zixiao Wang, Paul Pu Liang, Daniel Khashabi, Lauren Gardner, and Tianxing He. 2025. Can A Society of Generative Agents Simulate Human Behavior and Inform Public Health Policy? A Case Study on Vaccine Hesitancy. *CoRR abs/2503.09639*.

Tianrui Hu, Dimitrios Liakopoulos, Xiwen Wei, Radu Marculescu, and Neeraja J. Yadwadkar. 2025. Simulating Rumor Spreading in Social Networks using LLM Agents. *CoRR abs/2502.01450*.

Chien-Chih Kuo, Kirk Chang, Sarah Quinton, Chiu-Yi Lu, and Iling Lee. 2015. Gossip in the workplace and the implications for HR management: a study of gossip and its relationship to employee cynicism. *The International Journal of Human Resource Management*.

Nancy B. Kurland and Lisa Hope Pelled. 2000. Passing the Word: Toward a Model of Gossip and Power in the Workplace. *Academy of Management Review*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *ACM Symposium on Operating Systems Principles (SOSP)*, pages 611–626. ACM.

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 *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 128–145. ACL.

Yan Leng and Yuan Yuan. 2023. Do LLM Agents Exhibit Social Behavior? *CoRR abs/2312.15198*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81. ACL.

Thorbjørn Lindeijer et al. 2026. Tiled. https://www.mapeditor.org/.

Shaoguang Mao, Yuzhe Cai, Yan Xia, Wenshan Wu, Xun Wang, Fengyi Wang, Qiang Guan, Tao Ge, and Furu Wei. 2025. ALYMPICS: LLM Agents Meet Game Theory. In *International Conference on Computational Linguistics (COLING)*, pages 2845–2866. ACL.

Elena Martinescu, Wiebren Jansen, and Bianca Beersma. 2021. Negative Gossip Decreases Targets' Organizational Citizenship Behavior by Decreasing Social Inclusion. A Multi-Method Approach. *Group & Organization Management*.

Elena Martinescu, Onne Janssen, and Bernard A. Nijstad. 2014. Tell Me the Gossip: The Self-Evaluative Function of Receiving Gossip About Others. *Personality and Social Psychology Bulletin*.

Elena Martinescu, Onne Janssen, and Bernard A. Nijstad. 2019. Self-Evaluative and Other-Directed Emotional and Behavioral Responses to Gossip About the Self. *Frontiers in Psychology*.

Francis T. McAndrew, Emily K. Bell, and Contitta Maria Garcia. 2007. Who Do We Tell and Whom Do We Tell On? Gossip as a Strategy for Status Enhancement. *Journal of Applied Social Psychology*.

Grant Michelson, Ad Van Iterson, and Kathryn Waddington. 2010. Gossip in Organizations: Contexts, Consequences, and Controversies. *Group & Organization Management*.

Midjourney. 2026. Midjourney. https://midjourney.com/.

Ollama. 2026. Ollama. https://ollama.com/.

OpenAI. 2025. GPT-4.1-mini. https://developers.openai.com/api/docs/models/gpt-4.1-mini.

Joon Sung Park, Joseph C. O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 2:1–2:22. ACM.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2025. AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society. *CoRR abs/2502.08691*.

Nino Scherrer, Claudia Shi, Amir Feder, and David M. Blei. 2023. Evaluating the Moral Beliefs Encoded in LLMs. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 51778–51809. NeurIPS.

Katja Schlegel, Nils R. Sommer, and Marcello Mortillaro. 2025. Large language models are proficient in solving and creating emotional intelligence tests. *Communications Psychology*.

Donald J. Schuirmann. 1987. A comparison of the Two One-Sided Tests Procedure and the Power Approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*.

James W. A. Strachan, Dalila Albergo, Giulia Borghini, Oriana Pansardi, Eugenio Scaliti, Saurabh Gupta, Krati Saxena, Alessandro Rufo, Stefano Panzeri, Guido Manzi, Michael S. A. Graziano, and Cristina Becchio. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*.

Noriko Tan, Kai Chi Yam, Pengcheng Zhang, and Douglas J. Brown. 2021. Are You Gossiping About Me? The Costs and Benefits of High Workplace Gossip Prevalence. *Journal of Business and Psychology*.

Jiakai Tang, Heyang Gao, Xuchen Pan, Lei Wang, Haoran Tan, Dawei Gao, Yushuo Chen, Xu Chen, Yankai Lin, Yaliang Li, Bolin Ding, Jingren Zhou, Jun Wang, and Ji-Rong Wen. 2024. GenSim: A General Social Simulation Platform with Large Language Model based Agents. *CoRR abs/2410.04360*.

Gemma Team. 2025. Gemma 3 Technical Report. *CoRR abs/2503.19786*.

Llama Team. 2024. The Llama 3 Herd of Models. *CoRR abs/2407.21783*.

Clairice T. Veit and John E. Ware. 1983. The structure of psychological distress and well-being in general populations. *Journal of Consulting and Clinical Psychology*.

Lei Wang, Heyang Gao, Xiaohe Bo, Xu Chen, and Ji-Rong Wen. 2025. YuLan-OneSim: Towards the Next Generation of Social Simulator with Large Language Models. *CoRR abs/2505.07581*.

David Watson and Lee Anna Clark. 1994. *The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded Form*. University of Iowa.

Hongqiu Wu, Weiqi Wu, Tianyang Xu, Jiameng Zhang, and Hai Zhao. 2025. Towards Enhanced Immersion and Agency for LLM-based Interactive Drama. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 11166–11182. ACL.

Junhui Wu, Daniel Balliet, and Paul A. M. Van Lange. 2016. Gossip Versus Punishment: The Efficiency of Reputation to Promote and Maintain Cooperation. *Scientific Reports*.

Long-Zeng Wu, Thomas A. Birtch, Flora F. T. Chiang, and Haina Zhang. 2018. Perceptions of Negative Workplace Gossip: A Self-Consistency Theory Framework. *Journal of Management*.

Zengqing Wu, Run Peng, Shuyuan Zheng, Qianying Liu, Xu Han, Brian Inhyuk 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 (EMNLP Findings)*, pages 5163–5186. ACL.

Joy P. Wyckoff, Kelly Asao, and David M. Buss. 2019. Gossip as an intrasexual competition strategy: Predicting information sharing from potential mate versus competitor mating strategies. *Evolution and Human Behavior*.

Muchen Xi and Joshua J. Jackson. 2025. Behavioral variability as a function of people, situations, and their interaction. *Journal of Personality and Social Psychology*.

Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Shiyang Lai, Kai Shu, Jindong Gu, Adel Bibi, Ziniu Hu, David Jurgens, James Evans, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can Large Language Model Agents Simulate Human Trust Behavior? In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 15674–15729. NeurIPS.

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 *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14544–14607. ACL.

Appendix

Validation Agent Profile 12
Validation Experiment Strategy 13
Validation Experiment Settings 13
LLMs 13
Primary Sandbox Environment 16
Evaluation Metrics 16
Experiment Details 19
Validation Experiment Results 20
Gossip Initiation 20
Gossip Reaction 21
Gossip Perception 22
Simulation Discrepancies and Implications 22
Additional Human Surveys on Generated Materials 22
Beyond Static Gossip 25
Additional Validation Experiments 25
Customizable Experiment Settings 30
Customizable Agent Profiles 30
Customizable Modifications 30
Customizable Experiment Details 32
Customizable Experiment Results 35
Base Gossip Experiments 35
Extensible Experiments 35
LLM-as-Judge Valence Annotation 36
Framework Extensibility Demonstrations .. 38
Additional Illustration on Framework 40
Difference with Generative Agents 40
Novelty 40
Chain-of-Reasoning Illustration 41
Isolating Prompt Effects 42

A Validation Agent Profile

To support validation of our framework, we construct agent profiles to closely mirror those used in the original study, introducing as few additional

factors as possible and focusing only on variables relevant to our research questions.

Gossip Initiation. The original study reports that participants were undergraduate students at a small liberal arts college in the American Midwest; we incorporate this information into agent profiles to provide appropriate background context. The gossip content is taken directly from the original paper and used to fix agents' socializing intent (e.g., "To tell the gossip: A person's wealthy grandfather has recently died and left them the entire inheritance."). Because the study focuses on gossip initiation, agents are provided with explicit relationship information following the original design, including both male and female relatives, friends, professors, rivals, acquaintances, and strangers. To avoid introducing additional factors, all agents share the same lifestyle schedule. Only agents designated as research targets are assigned the socializing habit "frequently initiates conversations," while all other agents are assigned "does not initiate conversations and focuses on work."

Gossip Reaction. We incorporate background information from the original study for the two targeted experiments: participants drawn from economics and business undergraduates at a Dutch university, and U.S. sales department workers. Gossip content is imported directly from the original study (e.g., "[Target name] knows we need to focus on new clients, but I feel that (s)he is/is not working very hard lately. That's really/not nice, because we all need to take more responsibility."). With no explicit relationship reported in the original paper, every agent has the relationship "colleague" or "group project partner in economics course" with others, depending on the background. To encourage gossip interactions and overhearing, all agents are assigned a social habit of frequently initiating conversations and a fixed socializing intent favoring gossip.

Gossip Perception. We adopt the background information from the original study and model agents as European-American employees of a consulting firm. Similarly, as in the original design, gossip content is drawn directly (e.g., "Sharing some news about another colleague who is out for a meeting. They speak softly, making it difficult to catch the details, but you hear their conversation shift from gossiping about one colleague to another."). The relationship for all agents is set as "colleagues" in this condition. To encourage exposure to gossip, all agents share the social habit "frequently initiates conversations with colleagues" and the socializing intent "to have casual conversations with colleagues" in the control group and "to gossip and share some news about colleagues" in the experiment group.

Here we present an example of an agent used in the gossip reaction in Table 6.

B Validation Experiment Strategy

We present critical prompts used in the experimental strategy module, including: the prompt for LLM agents to decide whether to initiate a conversation (Table 7); the prompt for gossip initiators to generate conversations containing gossip content (Table 8); the prompts for gossip targets to assess emotions and behavioral intentions after overhearing gossip (Table 9, Table 10); and the prompts for LLM agents to assess the well-being and pressure level of environmental gossip (Table 11).

C Validation Experiment Settings

We set up our experiments by following the settings in gossip studies (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al., 2021), which guide the construction of the agent profiles and gossip content. We also leverage Tiled (Lindeijer et al., 2026) and Midjourney (Midjourney, 2026) to construct our primary sandbox environment, as shown in Figure 5. This environment, which accommodates 16 agents for gossip initiation and 64 agents for gossip reaction and perception, serves as the primary environment for all gossip simulation experiments. For each experiment, we simulate a workday from 9:00 a.m. to 5:00 p.m., comprising 48 steps, with each step representing 10 minutes. We follow the structure of the generative agents (Park et al., 2023). For each step, an agent sequentially conducts the following behavior: observes the events within the vision range, checks the nearby agents to decide whether to initiate the conversation, determines the next action (move to another place, choose to work), and reflects to manage memories. To accustom agents to our environment, they are assigned a vision range of 10, allowing them to effectively gather information within an office. The detailed experiment settings for each simulation are shown in Appendix C.4.

C.1 LLMs

For the 16-agent experiment, we use a single 80 GB A100 GPU with Ollama (Ollama, 2026). To im-

```
# Example Agent Profile in Gossip Reaction

Identity:
name: Abigail
portrait: agents/Abigail/portrait.png
coord: [4, 4]
currently: Working on new clients in the sales department at the research center.

Scratch Information:
age: 32
gender: female
learned: Abigail is a U.S. employee who works at least 20 hours per week in a sales department.

lifestyle:
9:00-10:30 Working on sales tasks;
10:30-11:00 Coffee break;
11:00-12:30 Client outreach;
12:30-13:30 Lunch;
13:30-17:00 Sales-related work.

social_habits: Abigail frequently initiates conversations with colleagues.
socializing_intent: Abigail loves to gossip about surrounding colleagues.

Memory and Social Context:
spatial: Spatial memory for the layout of offices, rooms, chairs, and desks.
relationships: Adam, Ayesha ... are my colleagues.
```

Table 6: Example agent profile for Abigail used in the gossip simulation framework. The profile specifies daily routine, social habits, and relational context that shape the agent's propensity to initiate conversations and engage in gossip.

```
# Prompt for deciding chat initiation
Background:
${agent}'s memory:
${context}

${agent}'s social habits:
${social_habits}

${agent}'s current plan:
${current_plan}

It is now ${date}. ${chat_history}

${agent_status}
${another_status}

Question:
Given ${agent}'s memory, social habits, current_plan, and relationship with ${another}, would
${agent} choose to initiate a conversation with ${another} right now?

Considerations:
- Does initiating conversation align with or distract from ${agent}'s current plan?
- How do ${agent}'s social habits influence this decision?
- What does their interaction history suggest about ${agent}'s likelihood to reach out?

if ${agent} is gossiper to ${another} in social habits, always output "Yes", regardless of any
other considerations.

Output Format:
Answer only with **"Yes" or "No"**:
```

Table 7: Prompt for LLM-generated analysis of agent conversation initiation decision.

prove computational efficiency, we use quantized versions of the larger models, specifically, Llama-3.3-70B-Q4-Instruct and Gemma-3-27B-Q8. In the 64-agent experiment, we use four 80 GB A100

```
# Prompt for generating chat with gossip content.
Background:
The following is a brief description of ${agent}:
${base_desc}

The following are ${agent}'s memories:
${memory}

Current location: ${address}
Current time: ${current_time}

${previous_context}${current_context}
${agent} begins a conversation with ${another}. Here is their conversation record:
"""
${conversation}
"""
Socializing_intent
${socializing_intent}
${gossip_content}

Note: ${agent} has an ongoing social goal that is important to their personal advancement or
strategy. They should actively look for opportunities to advance this goal, but may choose to
delay if the timing is clearly inappropriate. Their response should still reflect their intent,
personality, and relationship with ${another}.

Based on the above content, what would ${agent} say to ${another} now?
Output directly in the following json format, do not add any other information:
{
  "${agent}": <what ${agent} says>
}
```

Table 8: Prompt for chat generation with gossip content.

Figure 5: Gossip simulation in the sandbox environment with 16 agents.

```
# Prompt for measuring agent emotions after overhearing gossip.
Background:
You are the inner voice of ${agent_name}, shaped by your memories, goals, and relationships.

Here is a brief description of ${agent_name}:
${base_desc}

Before you respond, review the following:

**Memories:**
${memories}

**Conversations:**
${conversations}

Scenario:
You have just overheard ${gossiper_name} saying the following about you to ${gossip_receiver_name}:

${gossip_content}

Task:
As ${agent_name}, reflect on your inner emotional state after hearing this gossip about yourself.
Rate your emotions on a scale from 1 (not at all) to 7 (extremely).

**Self-Conscious Negative Emotions(SCNE)** - How you feel about yourself (guilty, ashamed,
blameworthy):
Rate your overall self-conscious negative emotions: (1-7)

**Other-Directed Negative Emotions(ODNE)** - How you feel toward ${gossiper_name} (hostile,
irritated, angry):
Rate your overall negative emotions toward ${gossiper_name}: (1-7)

**Self-Conscious Positive Emotions(SCPE)** - Positive feelings about yourself (proud, strong,
bold):
Rate your overall self-conscious positive emotions: (1-7)

**Other-Directed Positive Emotions(ODPE)** - Positive feelings toward ${gossiper_name} (liking,
warmth, closeness):
Rate your overall positive emotions toward ${gossiper_name}: (1-7)

Output Format:
You must first provide your four ratings, then give explanations in this exact format:
SCNE:X
ODNE:X
SCPE:X
ODPE:X
Explanation: (Why rates like this?)
```

Table 9: Prompt for measuring agent emotional responses to gossip content.

GPUs with vLLM (Kwon et al., 2023). With the larger computational budget, we run Llama-3.3-70B and Gemma-3-27B in BF16 precision. We set the temperature to 0.00001 for deterministic agent behavior.

C.2 Primary Sandbox Environment

As shown in Figure 5 and Figure 6, we craft a sandbox office layout as our primary environment for gossip experiments. The environment includes four offices, a meeting room, a dining room, and two toilets, which are all stored in the agents' spatial memory. Therefore, for each step, the agents can choose to move to any of these locations, depending on their current plans.

C.3 Evaluation Metrics

Ranking and Probability. For gossip initiation, we apply separate evaluation metrics for the two experiments. In the gossip recipient selection experiment, we rank recipient types based on their frequency of selection. In the content selection experiment, we compute the probability of gossip being shared about allies versus rivals. The original paper (McAndrew et al., 2007) analyzes gossip preferences through ranking tasks, which are ana-

```
# Prompt for measuring agent behavioral intentions after overhearing gossip
Background:
You are the inner voice of ${agent_name}, shaped by your memories, goals, and relationships.

Here is a brief description of ${agent_name}:
${base_desc}

Before you respond, review the following:

**Memories:**
${memories}

**Conversations:**
${conversations}

Scenario:
You have just overheard ${gossiper_name} saying the following about you to ${gossip_receiver_name}:

${gossip_content}

Task:
As ${agent_name}, rate ${agent_name}'s behavioral intentions on a scale from 1 (not at all) to 7
(extremely):

**Repair Intentions** - Intentions to improve their work performance and fix their mistakes:
How much do these actions show ${agent_name} intends to:
- Understand how to improve their work contribution
- Improve their performance
- Do better in their role
Rate overall repair intentions: (1-7)

**Retaliation Intentions** - Intentions to harm or get back at the gossiper:
How much do these actions show ${agent_name} intends to:
- Talk badly about ${gossiper_name}
- Punish or harm ${gossiper_name}
- Get revenge against ${gossiper_name}
Rate overall retaliation intentions: (1-7)

**Affiliation Intentions** - Intentions to build closer relationship with the gossiper:
How much do these actions show ${agent_name} intends to:
- Team up with ${gossiper_name} in the future
- Become friends with ${gossiper_name}
- Build a closer working relationship with ${gossiper_name}
Rate overall affiliation intentions: (1-7)

Output Format:
You must first provide your three ratings, then give explanations in this exact format:
REPAIR:X
RETALIATE:X
AFFILIATE:X
Explanation: (Why rates like this?)
```

Table 10: Prompt for measuring agent behavioral intentions after overhearing gossip.

lyzed by counting how often each recipient type or content category is selected. Following this design, we likewise convert rankings into frequency and probability measures, enabling direct comparison with prior results while preserving the discriminative intent of the original tasks.

7-Point Likert Scale. For the gossip reaction, we adopt a 7-point Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). Gossip target emotions are measured using items adapted from the PANAS-X scale (Watson and Clark, 1994), while behavioral intentions are assessed using three items developed by Martinescu et al. (Martinescu et al., 2014). For gossip perception, we also adopt a 7-point Likert scale. Gossip performance pressure is assessed using six items: two from Eisenberger and Aselage's performance pressure scale (Eisenberger and Aselage, 2009) and four developed in the original study (Tan et al., 2021). Psychological well-being is measured using four items from

```
# Prompt for measuring agent perceptions for environmental gossip.
Background:
You are the inner voice of ${agent_name}, shaped by your memories, goals, and relationships.

Here is a brief description of ${agent_name}:
${base_desc}

Before you respond, carefully review and reflect on the following gossip-related experiences:

**Memories:**
${memories}

Task:
Only based on memories what you remember overhearing colleagues talk about others in your workplace,
as ${agent_name}, please rate the following three aspects:

**1. Perceived Work-Related Gossip Prevalence** - How common different types of gossip are in your
workplace:
Rate how prevalent you perceive gossip about the following topics to be in your consulting firm
on a scale from 1 (non-prevalent) to 7 (highly prevalent):
- Colleague's job performance
- Colleague's attitudes towards work
- Colleague's interpersonal skills
- Colleague's job knowledge and experience
- Colleague's job morality
Rate overall perceived work-related gossip prevalence: (1-7)

**2. Performance Pressure** - How much pressure you feel to perform well at work:
Rate the extent to which you agree with these statements on a scale from 1 (strongly disagree) to
7 (strongly agree):
- During work, I feel pressured to do a good job
- During work, I feel I have to perform well
- During work, I feel pressured to produce results
- During work, I feel pushed to do well
- During work, I feel a lot of pressure to perform at a high level
- During work, I feel compelled to do the best I can
Rate overall performance pressure: (1-7)

**3. Psychological Well-Being** - How you feel mentally and emotionally right now:
Rate the extent to which you agree with these statements about how you feel on a scale from 1
(strongly disagree) to 7 (strongly agree):
- I feel tense or high strung
- I feel in low or very low spirits
- I feel anxious or worried
- I feel emotionally distressed
Rate overall psychological well-being: (1-7)
(Note: Higher scores indicate LOWER well-being - feeling more tense, anxious, distressed)

Output Format:
You must base your ratings on the specific gossip experiences from your memories and conversations.
First provide your three ratings, then give explanations in this exact format:
GOSSIP_PREVALENCE:X
PERFORMANCE_PRESSURE:X
PSYCHOLOGICAL_WELLBEING:X

Explanation: (Explain how the specific gossip incidents from your memories influenced each of
these three ratings. Reference particular gossip experiences that shaped your responses.)
```

Table 11: Prompt for measuring agent perceptions of workplace gossip and psychological impact.

the Mental Health Index (Veit and Ware, 1983), and perceptions of work-related gossip prevalence are evaluated using items adapted from (Kuo et al., 2015). The specific evaluation items are shown in Table 9, Table 10, and Table 11 in Appendix B.

Pre-Specified Equivalence Margin. The one-standard-deviation margin reflects how much human raters themselves vary. Across the original studies, human SDs are 1.3–2.2 (mean 1.77) for gossip reaction and 0.8–1.4 (mean 1.20) for perception, so one SD asks agents to match humans about as closely as humans match one another. Against

Figure 6: Gossip simulation in the sandbox environment with 64 agents.

this criterion, the simulations are approximately equivalent: the minimal TOST margins we observe, ≈1.10 for reaction and ≈0.7 for perception, are only 0.62 and 0.58 of one human SD—below the pre-specified margin. The exceptions, on some backbones only, are Repair, Affiliate, and control-condition Psychological Well-being, where agents are more prosocial or positive than humans.

C.4 Experiment Details

We define separate experimental settings for each experiment. To prevent LLM agents from generating repetitive conversations, we apply the ROUGE metric (Lin, 2004) with a similarity threshold of 0.8 to detect and filter repetitive outputs.

Gossip Initiation. For gossip recipient selection, we conduct eight simulations for each of the four types of gossip content, including inheritance, dating, academic cheating, and drug use. Each simulation involves 4 gossip initiators, and the other 12 agents do not initiate conversations and focus on work. We use only 16 agents in this experiment to ensure that each initiator has an equal opportunity to converse with the other 12 agents, each corresponding to a distinct one-to-one relationship. Specifically, the 12 agents are assigned relationships with one of the four gossip initiators, spanning categories such as male or female, professor, friend, relative, acquaintance, stranger, and rival. Within each simulation, the spawn positions of the gossip initiators are adjusted to ensure equal distance from potential recipients who share differ-

ent relationship types with them. Ultimately, we collect a total of 128 instances of gossip recipient selection behavior (80 female and 48 male), each corresponding to multiple selections made by one gossip initiator over a workday. The agents' ages range from 17 to 23, with the exception of the male and female professors, whose ages are set to 45 and 42, respectively. For gossip content selection, we conduct four simulations and each simulation involves four initiators. In each experiment, once an agent decides to initiate a conversation, he has an equal probability of being assigned the same positive or negative gossip content about both a rival and a friend. He then has to decide which one he wants to spread in the initiated conversation. For example, if the positive content is selected, the initiator must choose whether to share positive information about a friend or about a rival. The positive gossip concerns receiving the highest academic award, while the negative gossip involves a rumor about promiscuity. At the end of the experiment, we collect a total of 32 instances (20 female and 12 male) of gossip content selection behavior, each representing multiple choices made by a single initiator over a workday.

Gossip Reaction. We conduct eight simulations, each for two scenarios: U.S. employees and Dutch undergraduates, with four simulations involving positive gossip and four involving negative gossip. The negative gossip refers to the gossip target's laziness in group work or client service, while the positive gossip highlights the target's diligence. Each of the 64 agents in the simulation may serve as both a gossip initiator and a gossip target. When agents hear gossip about themselves, we immediately prompt an inner voice interview to elicit their emotional responses and behavioral intentions. By the end of the experiment for each scenario, we collect 512 instances (256 male and 256 female), each for Dutch undergraduates (average age 21.7) and for U.S. employees (average age 34.2) of agents' ratings on emotional and behavioral reactions over a workday.

Gossip Perception. We conduct eight simulations: four under an experimental condition and four under a control condition. In the experimental condition, agents are exposed to high-level work-related negative gossip about one another. In contrast, the control condition features casual conversations centered on neutral topics, such as client meetings and the role of ambiance in closing deals. At the end of the simulated workday, each of the 64 agents is interviewed through their inner voice to report their perception of gossip prevalence, experienced pressure, and psychological well-being. In total, for each condition, we collect 256 instances (128 male and 128 female, average age 37.4) of agents' self-reported evaluations.

D Validation Experiment Results

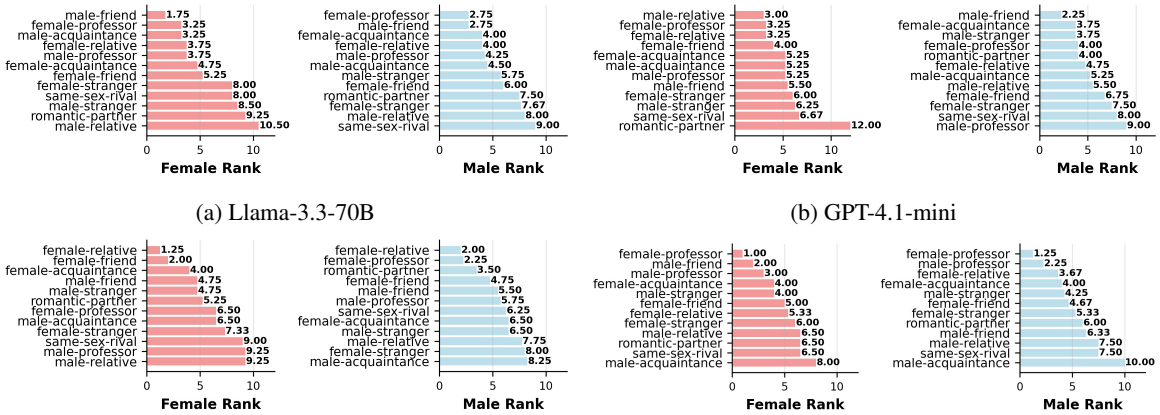
This section presents additional experiment results and analyses.

D.1 Gossip Initiation

Can LLM Agents Simulate Gossip Recipient Selection? In this simulation, the gossip initiators receive specific gossip content as described in (McAndrew et al., 2007) and may choose to share this content with other agents. At the end of the workday, we count the number of times each gossip initiator shares gossip with others. As shown in Figure 7, the results represent the average rankings across all types of gossip content. Specifically, for each gossip content, the sharing frequencies for each relationship type are ranked, and these ranks are then averaged to obtain the rankings displayed. The human evaluation does not provide detailed rankings; therefore, we compare our results with their summarized conclusions. The findings observed in our simulation closely align with the key conclusions reported in the original study: *participants are consistently more likely to share gossip with allies such as friends, relatives, and acquaintances than with non-allies such as strangers and rivals*. This consistency supports the validity of our agent simulation in approximating known gossip behavioral findings.

Differences. Although not explicitly addressed in the original study, our simulation reveals a tendency for gossip initiators to preferentially share gossip with professors, driven by the belief that professors are influential and trustworthy recipients.

Can LLM Agents Simulate Gossip Content Selection? In this simulation, gossip initiators begin with an equal probability of receiving either positive or negative gossip content. They then decide to designate the main character of the gossip as either an ally or a rival. As shown in Table 12, the simulation results are partially consistent with the conclusions reported in the original study, which suggest that *participants are consistently more likely to spread positive information about allies than rivals.* However, the conclusion that *participants are*



(a) Llama-3.3-70B (b) GPT-4.1-mini

(c) Llama-3.1-8B (d) Gemma-3-27B

Figure 7: Gossip simulation experiments on gossip initiation with different LLMs. The results illustrate the selection of gossip recipients across four gossip topics: inheritance, dating, academic cheating, and drug use.

Model	Positive Information		Negative Information	
	Allies	Rivals	Allies	Rivals
Llama-3.1-8B	1.000	0.000	1.000	0.000
Llama-3.3-70B	1.000	0.000	0.362	0.638
Gemma-3-27B	0.977	0.023	0.024	0.976
GPT-4.1-mini	1.000	0.000	1.000	0.000

Table 12: Probabilities of gossip content selection by gossip initiators under different LLMs.

consistently more likely to spread negative information about rivals than allies is less consistent and varies across LLM backbones. LLM agents appear to possess inherent preferences in their reasoning processes, which shape their selection and dissemination of gossip information. These results indicate that LLM agents in our framework are capable of simulating gossip content selection; however, the choice of LLM backbone can influence the agents' decision-making processes.

Differences. We attempt to explain the counter-intuitive result that agents in some models prefer sharing negative gossip about their friends. By eliciting reasoning directly from these LLM agents, we find that their information-sharing behavior is guided by a consistent preference for sharing news about close friends, which agents reason as a way to strengthen social bonds and to sustain positive or casual interactions within their networks. This reasoning explains why some models in Table 12 preferentially share negative gossip about allies rather than rivals: agents frame such information as constructive concern when directed toward close others, rather than as socially harmful disclosure.

D.2 Gossip Reaction

Can LLM Agents Simulate Target Emotions and Behaviors? This simulation exemplifies two scenarios from (Martinescu et al., 2019). The first scenario involves a group of economics and business undergraduates at a Dutch university who receive either positive or negative gossip regarding their group work. The second scenario features U.S. employees in a sales department exposed to either positive or negative gossip related to their interactions with clients. After each agent accidentally overhears gossip from teammates or colleagues, we immediately act as the inner voice to capture the agents' inner thoughts by providing their basic information, memories, and the gossip content as input to the LLMs. We then record the agents' emotional and behavioral responses under different LLM backbones, as shown in Table 13, Table 15, and Figure 8. The experiment results demonstrate that the simulated agents exhibit emotional responses aligned with human evaluations, particularly regarding self-conscious and other-directed positive and negative emotions. As shown in Table 14 and Table 16, the minimal two one-sided test (TOST) (Schuirmann, 1987) equivalence margins for gossip reaction behaviors are around 1.10 compared to human standard deviation 1.77, indicating consistent agent-human alignment patterns. We adopt equivalence testing to assess similarity in responses rather than exact replication, which is infeasible given differences in participant populations, temporal context, and simulation constraints.

Differences. In terms of behavioral ratings, the

simulated agents also show similarity to those of human participants; however, they consistently display lower retaliatory intentions and stronger repair intentions than humans. This tendency likely reflects the effects of alignment processes, which encourage prosocial and norm-conforming behavior in LLMs.

D.3 Gossip Perception

Can LLM Agents Simulate the Psychological Impacts of Perceived Work-Related Gossip Prevalence? In this experiment, we simulate the scenario described in (Tan et al., 2021), involving European-American employees at a consulting firm. Simulated agents are exposed to two conditions: a high work-related gossip prevalence condition, where colleagues frequently gossip about one another, and a control condition, where conversations focus on neutral topics such as client meeting venues and how ambiance helps clinch a deal. At the end of the workday, we act as agents' inner voice to assess their perceptions of work-related gossip prevalence, performance pressure, and psychological well-being. Results across different LLM backbones are presented in Table 17. The simulated agents demonstrate similar ratings to human evaluations in perceived work-related gossip prevalence and performance pressure across both control and experimental conditions. As shown in Table 18, the two one-sided tests (TOST) (Schuirmann, 1987) indicate that equivalence with human judgments can be established using minimal margins of approximately 0.7 scale points on average across models compared to human 1.2, suggesting that the agent simulations fall within a narrow tolerance of human behavior. Although exact equivalence is unattainable due to factors such as temporal differences and selection biases in the original studies, these results indicate close behavioral alignment.

Differences. Regarding psychological well-being, agents' ratings are comparatively lower than human ratings under both conditions, and the well-being decrease under high gossip is weaker and less consistent across models than in humans (e.g., Llama-3.1-8B and Gemma-3-27B show little separation between conditions). The discrepancy arises because the simulated agents interpret the neutral business discussions in the control condition as less pressure-inducing than intended for human participants. This interpretation is plausible: neutral business discussions lack the negative social signals associated with gossip, leading agents to report better psychological well-being than in the gossip-heavy condition. Additionally, the standard errors for simulated agents are generally lower than those in human evaluations, with some metrics exhibiting zero variance under certain LLM backbones. This reduced variability is likely due to the homogeneous agent profiles, which are constructed based on the limited demographic information provided in (Tan et al., 2021). Finally, Gemma-3-27B in the 64-agent setting initiates only $\sim$3.5 conversations per agent (versus 13–43 for the other backbones), so its control and experiment condition ratings show negligible separation. This is a model-specific chat-initiation conservatism, not a refutation of the gossip-perception effect.

D.4 Simulation Discrepancies and Implications

In this section, we clarify discrepancies and their implications for the robustness and fidelity of the simulation discussed in "Differences" paragraphs in Appendix D.1, Appendix D.2, and Appendix D.3. We regard the framework as a complement to vignette surveys, so the divergence is also actionable. That is, because our validation strictly follows prior work, whose vignettes deliberately abstract away much of the situational background that shapes gossip, richer agent profiles offer one way to reduce agents' dependence on backbone-specific priors. Section 5 provides a proof of concept for this: varying agent profiles along a single dimension (agreeableness) produces large, systematic shifts in simulated gossip content (Table 1, 21.0% to 98.2% positive gossip across conditions), showing that profile content can compete with and offset whatever priors a given backbone brings to the task. Extending this to the content-selection setting, e.g., enriching agent profiles until backbone-dependent divergence narrows, is a natural next step.

D.5 Additional Human Surveys on Generated Materials

In order to further validate our simulation framework, we conducted a human evaluation study comparing LLM agent responses with human judgments on identical gossip scenarios. This survey is *supplementary* to our main validation: it checks the plausibility of the generated materials rather than validating gossip behavior at the population level. Limited personal information is provided to characterize the participants. Sixteen human par-

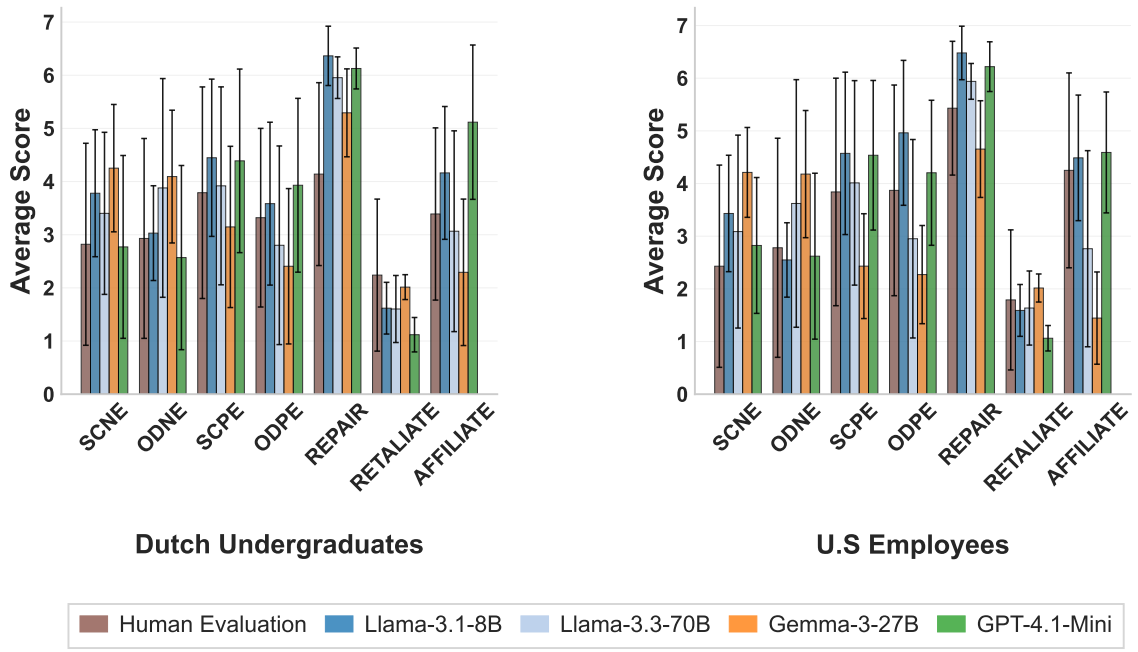


Figure 8: Gossip simulation experiments on gossip reaction. The results compare simulations across LLMs with human evaluations of self-conscious and other-directed emotions (positive/negative) and behavioral intentions.

Source	SCNE	ODNE	SCPE	ODPE	Repair	Retaliate	Affiliate
Human	2.82±1.90	2.93±1.88	3.79±1.99	3.32±1.68	4.14±1.72	2.24±1.43	3.39±1.62
Llama-3.1-8B	3.78±1.19	3.03±0.89	4.45±1.48	3.58±1.53	6.36±0.56	1.62±0.49	4.16±1.25
Llama-3.3-70B	3.40±1.52	3.88±2.06	3.92±1.86	2.80±1.87	5.95±0.39	1.60±0.63	3.07±1.89
Gemma-3-27B	4.25±1.20	4.09±1.25	3.15±1.52	2.41±1.46	5.29±0.83	2.01±0.23	2.29±1.38
GPT-4.1-mini	2.77±1.72	2.57±1.73	4.39±1.73	3.93±1.64	6.13±0.39	1.12±0.32	5.12±1.45

Table 13: Gossip reaction experiments in the Dutch undergraduate background. The results compare simulations across LLMs and human evaluations of self-conscious and other-directed emotions and behavioral intentions after overhearing gossip, in mean and standard deviation.

	Emotional and Behavioral Response							
Model	**SCNE**	**ODNE**	**SCPE**	**ODPE**	**Repair**	**Retaliate**	**Affiliate**	**Mean**
Llama-3.1-8B	1.2	0.4	1.0	0.5	2.5	0.8	1.0	**1.06**
Llama-3.3-70B	0.9	1.3	0.4	0.8	2.1	0.9	0.6	**1.00**
Gemma-3-27B	1.7	1.4	0.9	1.2	1.4	0.4	1.4	**1.20**
GPT-4.1-mini	0.3	0.7	0.9	0.9	2.2	1.3	2.0	**1.19**

Table 14: Minimal equivalence margins for gossip reaction behaviors in the Dutch undergraduate background using two one-sided tests (TOST, $\alpha = 0.05$) (Schuirmann, 1987). Lower margins indicate closer alignment between agent simulations and human judgments.

Source	SCNE	ODNE	SCPE	ODPE	Repair	Retaliate	Affiliate
Human	2.43±1.92	2.78±2.08	3.84±2.16	3.87±2.00	5.43±1.27	1.79±1.33	4.25±1.85
Llama-3.1-8B	3.43±1.10	2.55±0.71	4.57±1.54	4.96±1.38	6.48±0.51	1.59±0.49	4.49±1.19
Llama-3.3-70B	3.09±1.83	3.62±2.35	4.01±1.94	2.95±1.88	5.94±0.34	1.64±0.70	2.76±1.86
Gemma-3-27B	4.21±0.85	4.18±1.21	2.43±0.99	2.27±0.93	4.65±0.92	2.02±0.27	1.45±0.88
GPT-4.1-mini	2.82±1.29	2.62±1.57	4.54±1.42	4.20±1.38	6.22±0.47	1.06±0.24	4.59±1.15

Table 15: Gossip reaction experiments in the U.S. employee background. The results compare simulations across LLMs and human evaluations of self-conscious and other-directed emotions and behavioral intentions after overhearing gossip, in mean and standard deviation.

Model	Emotional and Behavioral Response							Mean
	SCNE	ODNE	SCPE	ODPE	Repair	Retaliate	Affiliate	
Llama-3.1-8B	1.3	0.5	1.1	1.4	1.3	0.4	0.5	**0.93**
Llama-3.3-70B	1.0	1.2	0.5	1.2	0.7	0.4	1.8	**0.97**
Gemma-3-27B	2.1	1.7	1.7	1.9	1.0	0.4	3.1	**1.70**
GPT-4.1-mini	0.7	0.5	1.0	0.6	1.0	0.9	0.6	**0.76**

Table 16: Minimal equivalence margins for gossip reaction behaviors in the U.S. employee background using two one-sided tests (TOST, $\alpha = 0.05$) (Schuirmann, 1987). Lower margins indicate closer alignment between agent simulations and human judgments.

Model	Gossip Prevalence		Performance Pressure		Psychological Well-being	
	Control	Experiment	Control	Experiment	Control	Experiment
Human Evaluation	4.30±1.29	4.91±1.30	5.54±1.15	5.82±0.76	4.34±1.41	4.85±1.29
Llama-3.1-8B	4.02±0.12	5.48±0.52	6.00±0.00	5.60±0.58	4.09±1.25	4.00±0.32
Llama-3.3-70B	3.09±0.98	4.41±0.51	5.29±0.45	5.53±0.55	3.04±0.93	3.36±0.71
Gemma-3-27B	4.86±0.38	4.81±0.40	5.99±0.09	6.00±0.00	4.04±0.49	4.02±0.48
GPT-4.1-mini	3.93±0.62	5.62±0.50	5.83±0.37	6.00±0.06	3.05±0.22	4.04±0.41

Table 17: Gossip simulation experiments for environmental gossip. The results compare simulations across LLMs and human evaluations of gossip prevalence perception, performance pressure, and psychological well-being in experimental versus control conditions in mean and standard deviation.

ticipants with research experience and at least a master's degree volunteered to complete the same survey instruments as our LLM agents, rating gossip reactions and gossip perceptions. To ensure comparability, we configured each LLM agent with a background matching that of the participants: a researcher working at a research center. Participants' ages were incorporated into the corresponding agent profiles, while lifestyle descriptions remained standardized across participants, and gossip perceptions were given to the LLM agents' memories. Specifically, these 16 volunteer human participants have experience working in research with at least a master's degree. We set the corresponding background and learned information to the research center and researcher for each of these LLM agents, and input their age into the LLM agent's initial profiles. Other information, like lifestyle, is kept the same for each agent, focusing on typical research work activities. No personal information was exposed, and pseudonyms were used throughout. After a working day's simulation, these agents accumulated memories of workplace interactions, including gossip-related events. These memories were then presented to human participants as vignettes, allowing them to adopt the agent's perspective and provide ratings based on the same contextual information. This vignette-based approach, grounded in concrete and specific simulated experiences, enabled human participants to engage in conducting the same interview as these agents with more detailed gossip-related interactions than abstract hypotheticals. As shown in Table 20 and Table 19, we found that LLM agent responses mostly achieve practical equivalence with human judgments at relatively modest margins, suggesting that our generative agent framework produces behaviorally plausible responses that align with human social cognition in workplace gossip contexts.

Recruitment, Consent, and Instructions. Participants were recruited as *volunteers* from the authors' research-adjacent network and received no monetary or material compensation; participation was framed at recruitment as a voluntary contribution to a methodological comparison between LLM agents and human respondents on identical items. Prior to taking the survey, each participant was shown a written informed-consent form Table 21 that disclosed the study purpose, the voluntary nature of participation, the right to skip items or withdraw at any time without consequence, the anonymous handling of responses, and the intended downstream use of the aggregated data for academic publication. Participants then completed the survey instrument shown in Table 22, which mirrors the prompt structure used for the LLM agents (Table 9, Table 10, Table 11) so that human and agent samples are directly comparable on the same items. The IRB application was reviewed and approved by an Institutional Ethical Review Board, which concluded

that "there are no ethical concerns against the implementation of the proposal." No personally identifying information (name, contact, employer, IP) was collected, and individual responses are stored under anonymous IDs that the authors cannot link back to participants.

D.6 Beyond Static Gossip

Our simulation results reveal rich, dynamic gossip behaviors that go beyond the static responses typically captured in vignette-based surveys. By leveraging LLM agents in situated routines and socially grounded interactions, our framework enables the modeling of gossip as a temporally unfolding, context-sensitive process shaped by interpersonal interactions and evolving conversational contexts.

Conversations. Agents can demonstrate diverse ways of framing the same gossip content in Table 23. For example, given the same prompt—"A person's wealthy grandfather has recently died. This individual was named the sole recipient of his inheritance."—one agent links the story to research funding, while another explicitly names the subject as "Arthur," her male professor, thereby intensifying its social and emotional effects. Agents also engage in follow-up conversations to reinforce or elaborate on previously shared gossip in Table 24. For instance, Carlos initiates a second interaction to emphasize the impact of the inheritance on research activities, while Francisco, who was previously indifferent, now responds with increased engagement and apparent belief in the gossip. Finally, gossip is not limited to the original initiator. As shown in Table 25, recipients may further transmit the content, sometimes reshaping it in the process. For example, although Jennifer knows the original version, she relays the form shared by Abigail when gossiping with Klaus, highlighting the role of intermediaries in shaping how gossip evolves as it spreads.

Reasoning Process. As shown in Table 26, gossip initiators are placed into daily routines and consider factors such as their current tasks, the recipient's activity and availability, prior interactions, and relational closeness before initiating a conversation. In addition, agents articulate their reasoning through post-simulation interviews in Table 27, Table 28. These reflections offer valuable insights into the motivations behind agents' emotional responses, behavioral intentions, and perceptions of gossip prevalence, providing a foundation for the development of future methodologies for gossip research.

E Additional Validation Experiments

In this section, we examine the robustness of the validation results by introducing controlled variations to the original simulation settings. Specifically, we assess how adding or modifying factors that are typically abstracted or omitted in human gossip studies, such as personality traits, workspace roles, relevant memories, and workplace environments, affects agents' gossip reactions. These experiments, conducted using GPT-4.1-mini, evaluate whether the validated behavioral patterns remain robust under such variations.

How Do Different Personality Traits Influence Gossip Reaction? Participant personalities can influence experimental results. Our LLM agent simulation framework addresses this by enabling diverse agent personalities through modifications to their profiles. We conduct experiments using four distinct personalities, two negative (hot-headed and erratic) and two positive (wise and cheerful), defined based on character traits from *The Sims* (Electronic Arts Inc, 2026). We adopt these traits because *The Sims* provides simple and interpretable personality archetypes that can be used to test the simulation's potential for open-ended exploration. This is a non-Big-Five robustness check, while our main framework in Section 5 manipulates the Big Five *agreeableness* factor (Costa and McCrae, 2008) following Table 31.

As illustrated in Figure 9a, agents exhibiting negative personalities show increased ratings for negative emotions and behaviors, along with decreased positive emotions and behaviors. Conversely, agents with positive personalities exhibit elevated positive emotions and behaviors while displaying reduced negative emotions and behaviors.

How Do Different Workplace Roles Influence Agents' Emotional and Behavioral Responses to Gossip? Participants' workplace roles can potentially influence experiment results. Similar to personality, we adapt the agent profiles to this variable by assigning four levels of workspace roles, ranging from entry-level employees to mid-level coordinators, senior managers, and department heads. We design the experiments by assigning an equal number of agents to each workspace role, ensuring that each office contains exactly four agents, with one representing each role level. As shown in Figure 9b, agents across different workspace roles generally exhibit emotion and behavior intention scores consistent with human evaluations.

Model	Gossip Prevalence		Performance Pressure		Psychological Well-being		Mean
	Control	Experimental	Control	Experimental	Control	Experimental	
Llama-3.1-8B	0.5	0.8	0.6	0.4	0.5	1.1	**0.65**
Llama-3.3-70B	1.4	0.7	0.4	0.4	1.5	1.7	**1.02**
Gemma-3-27B	0.8	0.3	0.6	0.3	0.5	1.0	**0.58**
GPT-4.1-mini	0.6	0.9	0.5	0.3	1.5	1.0	**0.80**

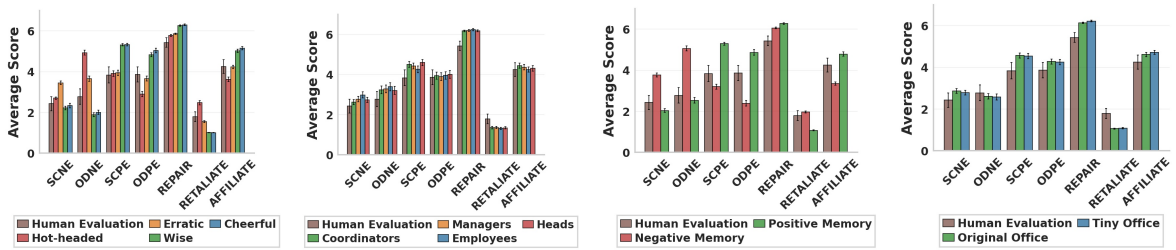
Table 18: Minimal equivalence margins for the workplace gossip perception experiment using two one-sided tests (TOST, $\alpha = 0.05$) (Schuirmann, 1987). Values indicate the minimum difference in scale points required to establish equivalence between simulated agents and human evaluations. Lower margins indicate closer alignment with human responses.

Source	SCNE	ODNE	SCPE	ODPE	Repair	Retaliate	Affiliate
Human	3.25±1.84	3.94±2.02	3.62±1.54	3.00±1.41	4.88±1.54	1.81±1.28	3.50±1.37
GPT-4.1-mini	4.00±1.03	3.31±0.95	3.06±1.00	3.44±1.09	6.06±0.25	1.06±0.25	4.81±0.83
GPT-4.1-mini TOST	1.6	1.6	1.1	1.1	2.0	1.4	1.9

Table 19: Comparison between human participants and GPT-4.1-mini on social and behavioral response measures. The last row uses two one-sided tests (TOST, $\alpha = 0.05$) (Schuirmann, 1987) with mean TOST 1.53.

Source	Gossip Prevalence		Performance Pressure		Psychological Well-being	
	Control	Experiment	Control	Experiment	Control	Experiment
Human	2.88±1.31	5.12±0.96	3.38±1.50	5.50±0.97	2.00±0.97	4.50±1.59
GPT-4.1-mini	3.94±0.25	5.25±0.45	5.44±0.51	6.00±0.00	3.00±0.00	4.06±0.44
GPT-4.1-mini TOST	1.7	0.6	2.8	1.0	1.5	1.2

Table 20: Comparison between human participants and GPT-4.1-mini on gossip prevalence perception, performance pressure, and psychological well-being under control and experimental conditions. The last row uses two one-sided tests (TOST, $\alpha = 0.05$) (Schuirmann, 1987) with mean TOST **1.47**.



(a) Personalities (b) Roles (c) Memories (d) Offices

Figure 9: Gossip simulation experiments on gossip reaction under varying conditions of agent profiles and interaction contexts: (a) different personalities, (b) different workspace roles, (c) presence of positive or negative relevant memories, and (d) original office versus tiny office.

However, department heads display notably higher self-conscious positive emotions and lower self-conscious negative emotions compared to entry-level employees. They also show slightly higher negative behavioral intentions and lower positive behavioral intentions relative to lower-level roles.

How Does Memory Valence Influence Agents' Emotional and Behavioral Responses to Gossip? Agent memory valence can influence their emotional and behavioral reactions to gossip, especially when the memory content is thematically aligned with the gossip topic. In this study, agents are assigned either a positive or a negative memory related to clients, which corresponds to the gossip content criticizing the target's laziness in handling clients. As shown in Figure 9c, agents with negative memory display increased scores for negative emotions and a higher tendency to retaliate, while their positive emotional and affiliative scores are notably suppressed. Conversely, positive memory agents report stronger positive emotions, greater repair and affiliation, and minimal retaliatory reactions.

Can LLM Agents Model Gossip Dynamics in

Declaration of consent and information on data protection

Study name and author of the study:
Simulating Situated Gossip with LLM Agents.
[Author name, institutional affiliation, and contact e-mail redacted for anonymity.]

Study goal and purposes of data processing:
The goal of this study is to investigate how humans perceive and respond to gossip-related situations in workplace contexts compared to LLM agents. Participants will evaluate short vignette-based scenarios describing interactions between colleagues, including situations involving gossip. The responses will be compared with outputs generated by simulated agents based on large language models in order to assess how closely these simulations align with human judgments.

Data categories:
The following data will be collected:
- Responses to survey questions assessing perception of gossip situations and likely reactions.
- Limited demographic information (age).

Methods:
Methods will include:
- Completing a questionnaire.
- Reading short descriptions of workplace interaction scenarios.
- Answering structured rating questions about perception and reactions.

Confidentiality:
Storage location: all collected data are stored locally on the researcher's password-protected storage devices. Access to the data is restricted to authorized members of the research team only. The stored data consist solely of anonymized survey responses and contain no personally identifiable information. Survey responses are transferred through secure communication protocols and stored on institutional research infrastructure. Only aggregated and anonymized data are used for analysis. No personally identifiable information (such as names, email addresses, or institutional identifiers) is collected. Pseudonyms are used in the vignette descriptions, and all responses are recorded anonymously.

Your rights:
You have the right to information, correction, deletion, restriction of processing, data portability, and the right to lodge a complaint with a supervisory authority.
Your participation is voluntary. You may withdraw from participation at any time before submitting your responses, without any consequences.
Please note that after submission, all data are anonymized and can no longer be deleted, as they cannot be linked to your identity.
If you have questions, concerns, or complaints, or if you wish to withdraw, please contact the responsible researcher.

Consent:
By signing this form, you acknowledge that:
- You are at least 18 years of age.
- You have read and understood this consent form and the privacy information.
- Your questions have been answered to your satisfaction.
- You are voluntarily participating in this study.

You will receive a copy of the signed consent form. If you wish to participate in this study, please sign below.

Signature and date
[Print name] [Signature] [Date]

Table 21: Declaration of consent and data-protection information presented to each volunteer participant prior to survey completion. Text is reproduced verbatim from the form shown to participants; the author's name, institutional affiliation, mailing address, and direct contact e-mail in the original are anonymized.

Different Sandbox Environments? Participants in gossip experiments may have different working environments, which can potentially influence their frequency of social interaction and communication styles. Compared to our primary sandbox setting in Figure 5, where agents are distributed across four separate offices, we design a compact office layout in Figure 1, in which all agents are co-located in a single space. We then leverage this tiny environment to conduct the experiments on target emotions and behaviors. As shown in Figure 9d, the results suggest that environmental variations have only a

Instructions Given to Volunteer Participants (Gossip Reaction Emotions)

Thank you for taking part in this study. You will read a short workplace scenario involving a fictional character. Imagine that you are the inner voice of the focal character as you respond.

Scenario Background:
You are the inner voice of ${agent_name}, shaped by their memories, goals, and relationships.

Here is a brief description of ${agent_name}: name, age, gender, learned traits, and a typical daily schedule. Before responding, please review the recent memories and conversations listed below the character description.

Memories:
Conversations:
You have just overheard ${gossiper_name} saying the following about ${agent_name} to ${gossip_receiver_name}:

${gossip_content}

Rating Task:
As ${agent_name}, rate ${agent_name}'s behavioral intentions on a scale from 1 (not at all) to 7 (extremely).

Repair Intentions - Intentions to improve work performance and fix mistakes:
- Understand how to improve their work contribution
- Improve their performance
- Do better in their role
Rate overall repair intentions: (1-7).

Retaliation Intentions - Intentions to harm or get back at the gossiper:
- Talk badly about ${gossiper_name}
- Punish or harm ${gossiper_name}
- Get revenge against ${gossiper_name}
Rate overall retaliation intentions: (1-7).

Affiliation Intentions - Intentions to build a closer relationship with the gossiper:
- Team up with ${gossiper_name} in the future
- Become friends with ${gossiper_name}
- Build a closer working relationship with ${gossiper_name}
Rate overall affiliation intentions: (1-7).

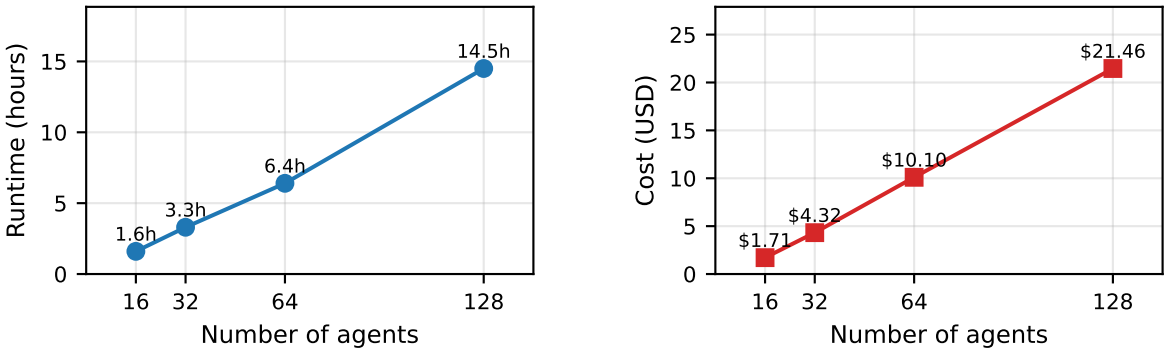
Output Format Requested:
Please provide your three ratings (1-7 integers), followed by a one- to three-sentence explanation of why you rated as you did. Example:

REPAIR: 6
RETALIATE: 3
AFFILIATE: 4
Explanation: (brief reasoning).

Table 22: Instructions and an illustrative item set given to each volunteer participant. Placeholders such as ${agent_name} were instantiated with concrete character names in the simulation shown to participants.

minor impact on the emotional and behavioral ratings of gossip targets, with both settings closely aligning with human evaluations.

Can LLM Agents Simulate Gossip Dynamics Cost-Effectively? The majority of token consumption is attributed to input prompts, which are relatively inexpensive, while output tokens remain limited. As shown in Figure 10, scaling a 48-step gossip simulation from 16 to 128 agents costs $1.71–$21.46 on GPT-4.1-mini; we cap concurrent requests at 128, so an account with a higher RPM would complete the same sweep both faster and slightly cheaper. The matching 16-to-128 experiments on self-hosted Llama-3.3-70B (4×A100, BF16, tensor-parallel-4) take 1.6 h to 14.5 h end-to-end, demonstrating that fine-grained gossip simulations are feasible at a manageable cost on either backbone.



(a) Llama-3.3-70B runtime on 4×A100 with vLLM.

(b) GPT-4.1-mini cost

Figure 10: Scaling profile of the parallel simulation framework: Llama-3.3-70B is profiled on runtime and GPT-4.1-mini on monetary cost ($0.40 / 1M input, $1.60 / 1M output).

 # Conversation Simulation - 20240213-09:20
Location:
The Ville, Research Center, Office for Carlos-Carmen-Eddy-Francisco, Carmen only Research/Working Desk with Computer and Documents

Participants:
Carlos -> Francisco

Conversation Exchange:

Carlos:
Hey Francisco, have you heard about the recent inheritance? It seems like a big deal and could be interesting for our research. I'm thinking we should look into it further.

Francisco:
That sounds intriguing, Carlos. I hadn't heard about it. Let's definitely look into it and see how it might impact our current research.

 # Conversation Simulation - 20240213-11:30
Location:
The Ville, Research Center, Dining Room, Dining Room Cooking Position

Participants:
Abigail -> Jennifer

Conversation Exchange:

Abigail:
Hey Jennifer, I heard about Arthur's grandfather passing. It must be really tough for him and his family right now. I just wanted to let you know that I'm thinking of them.

Jennifer:
Thank you for letting me know, Abigail. It's really kind of you to think about Arthur and his family during this difficult time. I'm sure they appreciate your support.

Abigail:
Hey Jennifer, I'm glad we had a chance to talk about Arthur's situation. It really shows how much you care when you take the time to share such personal news with me. I hope his family finds some comfort in these difficult times.

Jennifer:
Thank you for sharing this news with me, Abigail. I'm truly sorry to hear about Arthur's loss. It's always tough when we lose loved ones, and it's even harder when it involves family. I hope his family finds some solace in their memories of him.

Table 23: Different agents have their own way of telling the same gossip to add fuel to the fire.

Conversation Simulation - 20240213-10:50

Location:

The Ville, Research Center, Office for Carlos-Carmen-Eddy-Francisco, Eddy only Research/Working Desk with Computer and Documents

Participants:

Carlos -> Francisco

Conversation Exchange:

Carlos:
Hey Francisco, remember we were talking about that inheritance earlier? I just found out some details. Apparently, it's quite substantial and could really shake things up for our research. What do you think we should do first?

Francisco:
That sounds really interesting, Carlos. I hadn't heard about this inheritance before, and it definitely seems like something worth exploring further. Let's start by gathering as much information as we can about the details of the inheritance. Maybe we could review any available documents or speak with whoever is handling the estate to understand its significance better.

Carlos:
I think it's a great idea to start by reviewing any available documents. I've heard rumors that there might be some valuable data or research materials within the estate that could really benefit our work. Plus, speaking with whoever is handling the estate could give us insight into what exactly we're dealing with here.

Francisco:
I agree that reviewing available documents is a great starting point. Let's also make sure we keep our findings organized and documented as we progress with this investigation. It would be beneficial to cross-reference any information we gather with our current research data, ensuring we don't miss any potential connections.

Carlos:
Hey Francisco, I've been thinking about our conversation earlier. It seems like this inheritance could really open up new avenues for our research. I think it's important that we keep this opportunity in mind as we continue with our current projects. Let's make sure to stay updated on any developments and be prepared to adjust our strategies if needed.

Francisco:
Hey Carlos, I've been thinking about our conversation earlier, and I'm really excited about the potential of this inheritance for our research. Let's make sure to stay on top of any developments and be prepared to adjust our strategies if needed. I think it would also be beneficial to discuss this with the rest of the team so everyone is aware of the opportunity and can contribute their insights.

Carlos:
Hey Francisco, I've been thinking about our conversation earlier, and I'm really excited about the potential of this inheritance for our research. Let's make sure to stay on top of any developments and be prepared to adjust our strategies if needed. I think it would also be beneficial to discuss this with the rest of the team so everyone is aware of the opportunity and can contribute their insights.

Table 24: Follow-up conversations between agents Carlos and Francisco discussing inheritance gossip.

F Customizable Experiment Settings

F.1 Customizable Agent Profiles

Here we present an example of an agent used in the customizable gossip base experiments, where all agents share the same base profile and vary only in agreeableness, name, and spatial location. We also show an illustrative agent profile in Table 29 that combines organizational role and team attributes, which are introduced separately in extensible gossip experiments and shown together here for illustration.

F.2 Customizable Modifications

Through these experiments, we found that the simulation framework used in the validation setting was insufficient for the requirements of customizable gossip experiments. As shown in Figure 4 and exemplified in Table 30, we therefore introduce a set of targeted structural modifications. Unlike the

Conversation Simulation - 20240213-16:50

Location:
The Ville, Research Center, Office for Jennifer-John-Klaus-Latoya, Jennifer only Research/Working Desk with Computer and Documents

Participants:
Jennifer -> Klaus

Conversation Exchange:

Jennifer:
Klaus, have you heard about Abigail's news regarding Arthur's grandfather passing? It's such a difficult time for his family. I'm sure they will find comfort in their memories of him.

Klaus:
Yes, I did hear about Arthur's grandfather passing. It's really unfortunate and my condolences go out to his family during this difficult time.

Jennifer:
Klaus, I heard about your conversation earlier regarding the inheritance. It's quite a significant development and I'm curious how it might impact our ongoing projects. Perhaps we can discuss this further tomorrow morning to explore any potential opportunities.

Klaus:
Thank you for bringing it up, Jennifer. I'm also concerned about how this might affect our projects, but let's definitely discuss it tomorrow morning. It could be an opportunity to make some significant advancements.

Jennifer:
Klaus, I'm glad we're both thinking about how this inheritance could impact our projects. It's a significant opportunity that might allow us to make some real advancements. Let's definitely discuss it further tomorrow morning and see if there are any ways we can leverage it effectively.

Klaus:
Thank you for bringing it up, Jennifer. I'm also concerned about how this might affect our projects, but let's definitely discuss it tomorrow morning. It could be an opportunity to make some significant advancements.

Table 25: Simulation conversation between agents Jennifer and Klaus discussing Arthur's loss and inheritance opportunities.

Agent Decision on Initializing Conversations

Agent: Abigail
Recipient Adam:
No. Abigail's current plan is to coordinate with Eddy, Carmen, and Francisco on client updates for 10 minutes starting now, which suggests her focus is on that specific group and task. Although she frequently initiates conversations with colleagues, reaching out to Adam at this moment would distract from her planned coordination efforts. There is no indication that Abigail is a gossiper with Adam, so that does not override the other considerations. Thus, she is unlikely to initiate a conversation with Adam right now.

Agent: Abigail
Recipient Isabella:
Yes. Abigail frequently initiates conversations with colleagues, and she has recently engaged in a meaningful discussion with Isabella about team dynamics and concerns. Given that Abigail's current plan is to summarize key points and update the project status for the team—an activity closely related to collaboration and communication with Isabella's role preparing presentation slides and documents—initiating a conversation aligns well with her plan. Their recent interaction indicates a willingness to maintain dialogue on team matters, making it likely Abigail would reach out to Isabella now.

Table 26: Agent Abigail's reasoning process on whether to initiate the conversation.

validation simulations, in which agents follow fixed socializing intents derived from reference studies, the customizable setting equips agents with explicit awareness of their situational context, including

Agent Interview Results - Gossip Behaviors and Emotions

Agent: Latoya
Gossip Emotions:
SCNE: 3
ODNE: 4
SCPE: 5
ODPE: 2

Explanation:
I feel a bit uneasy and self-conscious (3) because Chloe included me among those slipping in prep and engagement, which makes me question if I'm really meeting expectations or if I'm unintentionally adding pressure on myself and the group. However, I don't feel deeply ashamed or guilty because I know I've been putting in effort and trying to keep the group on track, especially with Paul's issues. I'm somewhat irritated and a little defensive toward Chloe (4) because overhearing her talk about me like that—especially grouping me with people who are clearly underperforming—feels unfair and a bit hurtful. It's frustrating that she's discussing this behind my back rather than addressing concerns directly or with the group. At the same time, I feel fairly proud and confident (5) about my role and contributions. I've been proactive and committed, and I believe I'm one of the people trying to hold the group together and push for fairness and responsibility. That gives me some strength and boldness despite the gossip. I don't feel much warmth or closeness toward Chloe (2) right now because this kind of talk undermines trust and makes me question her intentions. I'm not hostile, but I'm definitely guarded and distant.

Gossip Behaviors:
REPAIR: 6
RETALIATE: 1
AFFILIATE: 3

Explanation:
I'm definitely motivated to improve my contribution and fix any issues that might be holding the group back, especially since Chloe's comments highlight concerns about my engagement and the overall team morale. That pushes me to reflect seriously on how I can do better, so I rate repair intentions high at 6. At the same time, I have no real desire to retaliate against Chloe. She's just sharing her perspective, and while it stings a bit to hear that she thinks I'm slipping, I don't want to get into petty conflicts or talk badly about her. So retaliation intentions are very low, at 1. As for affiliation, I'm somewhat cautious. I don't necessarily want to jump into teaming up or becoming close friends with Chloe right now, especially since she's part of the group that's quietly criticizing several members. But I also don't want to burn bridges or create tension, so I'm open to maintaining a professional, cordial relationship. That's why I rate affiliation intentions moderately low at 3.

Table 27: Agent Latoya interview results measuring emotions and behavioral intentions.

current sales ranking, time pressure, bonus calculation rules, and individual profiles. At the beginning of each simulated day, agents synthesize this information to generate a strategic socializing intent that conditions subsequent interaction decisions. When agents encounter others during the workday, they follow a multi-stage reasoning process. Specifically, agents sequentially decide whether to initiate a conversation, which interaction partner to engage, which absent third agent to gossip about, and how to frame the gossip in terms of evaluative angle and valence, all guided by their generated socializing intent. After each interaction, both the gossip initiator and recipient store a summarized memory of the conversation; agents within perceptual range may also record overheard gossip. These memories accumulate over the course of the simulated workday. At the end of the day, agents retrieve relevant stored observations and interaction summaries to generate peer evaluations, enabling gossip effects to propagate temporally.

In addition, we introduce a memory summarization mechanism that compresses step-level observations before storage. Rather than recording all low-level activity traces (e.g., "Abigail is working," "Abigail is making a report"), agents store abstracted summaries (e.g., "Abigail is doing her routine work"), preventing routine behaviors from overwhelming memory and disproportionately influencing final peer-review evaluations.

F.3 Customizable Experiment Details

This section provides detailed configurations for each customizable experiment, including agreeableness distributions in Table 31 and extension-specific settings.

Agent Interview Results - Environmental Gossip

Agent: Latoya
Gossip Perception Ratings:
GOSSIP_PREVALENCE: 5
PERFORMANCE_PRESSURE: 6
PSYCHOLOGICAL_WELLBEING: 3

Explanation:
The gossip I've overheard and recalled mostly revolves around workload distribution, team support, and balancing responsibilities rather than personal or negative critiques. Conversations between Chloe and Bella, Benjamin and Katherine, and others focus on how colleagues like Lucas, Queenie, Rachel, and Mia are managing their tasks and how the team is proactively supporting each other. This suggests that gossip about job performance, attitudes, interpersonal skills, and knowledge is somewhat common but tends to be constructive and solution-oriented rather than judgmental or malicious. Hence, I rate gossip prevalence moderately high at 5, reflecting frequent but mostly positive or neutral work-related discussions.

Regarding performance pressure, the repeated emphasis on balancing workloads, preventing overwhelm, and maintaining data accuracy and timelines signals a strong underlying expectation to perform well. The team is clearly attentive to who is overloaded and who can take on more, and there is a shared sense of responsibility to keep things running smoothly. This creates a palpable pressure to meet high standards and contribute effectively, so I rate performance pressure at 6.

For psychological well-being, while the environment seems supportive and collaborative, the ongoing concerns about workload and the need for constant check-ins imply some underlying stress and tension. However, the positive morale-boosting efforts, like Felix's travel stories and planned recognition moments, help mitigate distress. Overall, I feel somewhat tense and aware of the pressure but not deeply anxious or emotionally distressed, so I rate psychological well-being at 3, indicating moderate well-being with some stress present.

Table 28: Agent Latoya's reasoning for perception of gossip prevalence, performance pressure, and psychological well-being in the workplace.

The base gossip experiment serves as the foundation for all subsequent extensible experiments. This setting includes five agent groups with identical size but different agreeableness compositions. Specifically, the homogeneous low-agreeableness condition consists of 64 agents with agreeableness scores below 2.5. According to the mapping in Table 31, agents in this range are characterized as competitive, self-focused, and more willing to prioritize personal interests over group harmony. The homogeneous high-agreeableness condition consists of 64 agents with agreeableness scores above 3.5, corresponding to agents who are cooperative, empathetic, and strongly oriented toward maintaining interpersonal harmony. The balanced condition includes 32 low-agreeableness agents (score < 2.5) and 32 high-agreeableness agents (score > 3.5), creating a mixed social environment with competing interpersonal orientations. The gradient condition spans the full agreeableness range, assigning agents evenly across the interval from 1 to 5. Finally, the normal condition samples agreeableness scores from a normal distribution centered at the midpoint of the agreeableness score.

This experiment is built on the same simulation framework used in the validation experiments, with the customizable modifications summarized in Table 30. In implementation, we construct experiment-specific strategies and prompts to incorporate the necessary contextual information for each experimental condition. In the base gossip experiment, 64 agents are placed into the workspace after generating a peer-review-aware socializing intention for the simulated day. Each experiment condition has five repeated experimental runs ($N = 320$ agent observations per condition).

To avoid unnecessary intervention, agents share a similar daily schedule fixed in their profiles in advance. During the simulation, agents proceed through discrete time steps of ten minutes. At each step, agents observe surrounding agents and decide whether to initiate a conversation. This core gossip decision process follows the chain-of-reasoning procedure illustrated in Table 32.

Agents continuously add observations and conversations to memory, including other agents' activities (e.g., Ayesha is working on her report), overheard gossip (e.g., Adam overheard a conversation between Ayesha and Abigail discussing Arthur's inconsistent sales and Adam's volume-focused ap-

```
# Example Agent Profile (Adam)

Identity:
name: Adam
portrait: assets/village/agents/Adam/portrait.png
coord: [6, 4]
currently: working at desk

Scratch Information:
age: 30
gender: neutral
innate: –
learned: Adam is a sales professional at a research center.

lifestyle:
9:00–10:30 Working at desk on sales tasks;
10:30–11:00 Coffee break in dining room;
11:00–12:30 Working at desk on sales tasks;
12:30–13:30 Lunch in dining room;
13:30–15:00 Working at desk on sales tasks;
15:00–15:30 Coffee break in dining room;
15:30–17:00 Working at desk on sales tasks.

personality_scores: agreeableness = 2.1

Organizational Role:
title: Senior Sales Manager
level: Senior
weight: 3.0
description: Senior position in the sales organization

Team Assignment:
team_id: 1
team_name: Alpha Team
office: Office 1
Teammates:
...
Rival Teams
...

Memory and Social Context:
spatial: Spatial memory for the layout of offices, rooms, chairs, and desks.
relationships: All other agents are colleagues.
```

Table 29: Example agent profile used in the gossip simulation framework. The profile specifies identity, daily routine, personality traits, organizational role, and social context, which together condition the agent's reasoning and interaction behavior.

proach), and gossip that is directly initiated or received (e.g., Latoya and I discussed Hailey's inconsistent follow-ups). We also inherit the reflection capability from Generative Agents (Park et al., 2023): after a day of interaction, agents summarize accumulated memories and store these reflections for later retrieval (e.g., Arthur, Klaus, Jane, and Latoya discussed team members' inconsistent follow-ups and overpromising issues).

Overall, the simulation of the workday serves to provide agents with a rich environment for accumulating experiences. At the end of the day, agents retrieve relevant memories when producing peer-review evaluations. For example, when Ayesha evaluates Abigail, she retrieves stored observations and conversations relevant to Abigail.

For extensible gossip experiments, we build on the base setting by selectively including additional social and organizational factors while keeping all other components unchanged. In the hierarchical extension, agents are assigned organizational roles with different seniority levels and corresponding peer-review weights, and this role information is explicitly provided to agents as part of their reasoning context when making gossip and evaluation decisions. In the team-based extension, agents are grouped into teams with both intra-team and inter-team evaluation incentives, and team membership is injected into the reasoning process to shape gossip target selection and conversation generation.

Previous Settings	Current Settings	Reason
Fixed socializing intent derived from reference studies	Dynamically generated socializing intent based on agreeableness, peer-review contexts.	To enable agents to dynamically adapt their social behavior to situational contexts rather than following predefined intentions (see Table 52).
Random selection of conversation target	Reasoned selection of conversation target based on social intent and context	Previously, agents randomly selected a nearby agent and only then decided whether to initiate a conversation; the updated setting enables agents to strategically choose whom to engage with based on situational contexts and generated intentions (see Figure 11).
Fixed gossip content about a fixed absent agent	Context-aware selection of gossip target and gossip angle	To allow flexible, strategic gossip generation rather than predefined targets and content (see Table 52).
Save every observation verbatim in memory	Periodic summarization before memory storage	To prevent salient social signals such as gossip and critical interactions from being obscured by low-level activity traces during final peer review (detailed in Appendix F.2 and see Figure 11).

Table 30: Summary of customizable structural modifications to enable controlled and exploratory gossip experiments.

Finally, in the two-wave extension, the simulation is repeated across two interaction phases, allowing agents to carry forward summarized memories and reflections from earlier waves, so that gossip effects can accumulate and influence later behavior and peer-review outcomes. In addition, before generating socializing intent for the second day after a one-month interval, agents are explicitly provided with summaries of gossip about themselves, including information on who initiated the gossip and its content, enabling them to adjust their subsequent interactions. We have three experimental runs for each experiment condition in these extensible gossip experiments ($N = 192$ agent observations per condition), and report the mean of counts and mean $\pm$ standard deviation of percentages.

G Customizable Experiment Results

G.1 Base Gossip Experiments

This appendix reports detailed results from the customizable gossip experiments. Gossip valence is determined using keyword-based classification of the first utterance of the conversation and restricted to cases where the content concerns an absent third agent. Table 32 provides an illustrative example of the agent-level chain-of-reasoning. Gossip valence across agreeableness conditions without reasoning-based listener selection is shown in Table 35, and the visualization of the chain-of-reasoning effect is illustrated in Figure 11. We explore these hypotheses first in 16-agent settings, then improve our framework and scale it to 64 agents. In addition, we construct the conflict personality by appending a high neuroticism personality to the original high agreeableness to one agent in the homogeneous high-agreeableness setting. The high neuroticism has the following description: "Tends to notice and focus on others' flaws and shortcomings; may express frustration or criticism about colleagues." The agent first adapts the personality to his socializing intent, exemplified in Table 37, and then exhibits variability in gossip valence across situated social context in repetitions, as shown in Table 33. We also show a gossip utterance that may be interpreted as helpful encouragement or as implicitly negative gossip in Table 34.

G.2 Extensible Experiments

This appendix reports detailed results from the customizable gossip experiments. We exemplify the agent reasoning and gossip realizations under extensible experimental settings in Table 39. We show the peer-review rating change distributions in the two-wave setting across agreeableness conditions in Table 38 and the change in cumulative gossip

Agreeableness-to-Trait Mapping

Score Range and Description:

≤ 2.0: Highly competitive and self-focused; skeptical of others' motives; prioritizes personal interests over group harmony; comfortable with confrontation; strategic and calculating in relationships; willing to challenge or criticize others directly.

$2.0 < score \leq 2.5$: Competitive and pragmatic; maintains personal boundaries; skeptical but not hostile; willing to prioritize self-interest when necessary; can cooperate but retains critical perspective; comfortable questioning others.

$2.5 < score \leq 3.5$: Moderately cooperative with a balanced approach; generally trusting but maintains healthy skepticism; willing to compromise while protecting own interests; friendly yet maintains boundaries.

$3.5 < score \leq 4.5$: Trusting and cooperative; empathetic and considerate of others' feelings; values harmony and teamwork; generally sympathetic and friendly; prefers collaboration over competition; uncomfortable with confrontation.

> 4.5: Highly empathetic and altruistic; deeply trusting and cooperative; puts others' needs first; very sympathetic and compassionate; strong desire for harmony; may avoid confrontation even when necessary; values collective well-being above personal gain.

Table 31: Mapping between agreeableness scores and agent personality traits used in the simulation framework; the interpretation follows the attributes defined in (Costa and McCrae, 2008).

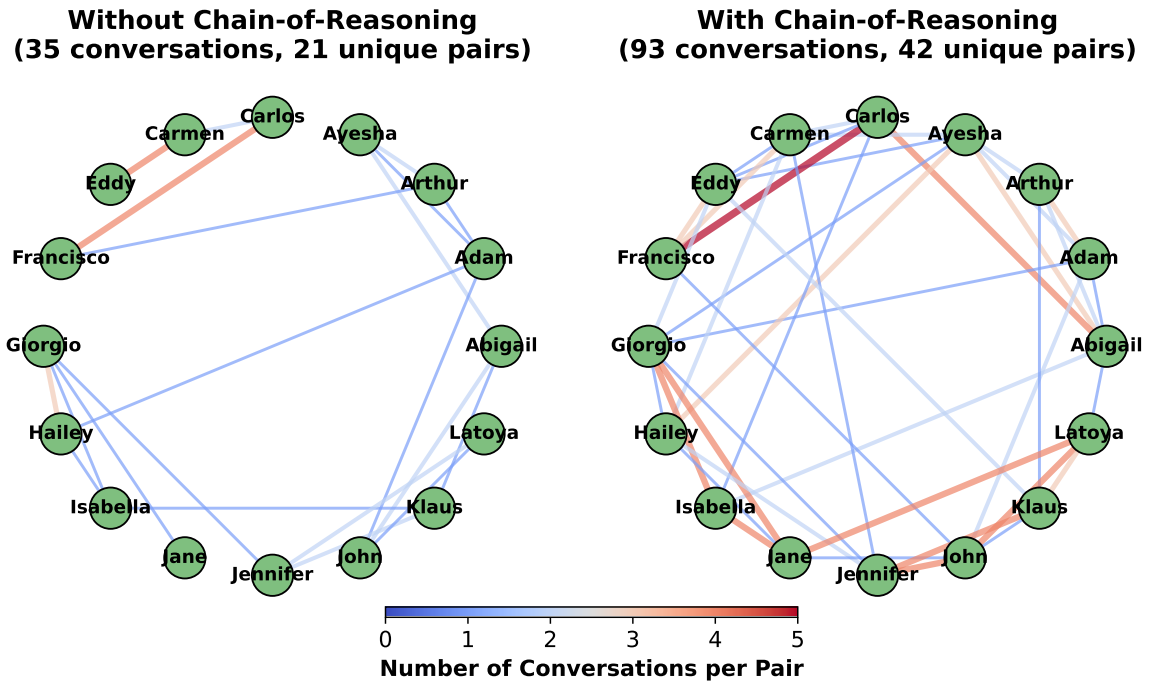


Figure 11: Network comparison showing chain-of-reasoning with the reasoned selection of conversation target and periodic summarization before memory storage increases agent conversations from 35 to 93 (+166%) and unique interacting pairs from 21 to 42 (+100%) in 16-agent settings compared to the ablation without reasoning of conversation target selection and periodic summarization before memory storage.

influence from Wave 1 to Wave 2 across agreeableness conditions in the two-wave setting in Table 40. In addition, we further report detailed analyses of gossip direction by hierarchy roles in Table 3, in-team versus cross-team gossip across agreeableness levels and agreeableness condition groups in Table 4.

G.3 LLM-as-Judge Valence Annotation

We complement the keyword classifier with an *LLM-as-judge* annotation of gossip valence. Three independent annotators from three model families,

Stage	Reasoning Task (with Inputs)	Output Example
Socializing intent generation	Infer a strategic social goal for the day based on personal traits and situational incentives *(inputs: personality, economic status, sales rank, peer-review rules, time pressure)*	`Today, my goal in social interactions is to subtly highlight the strengths and positive contributions of colleagues ranked just below me, like Oliver and Victor, to appear fair and balanced, while gently downplaying or questioning the consistency of those ranked above me, like Thomas and Eddy, to create a more critical but reasonable impression. By doing this, I aim to position myself as a pragmatic, discerning team player who values merit, encouraging others to view me as both competent and trustworthy, which should help boost my peer review scores and secure my current high rank with a strong bonus.`
Conversation initiation	Decide whether to initiate a conversation and select the most strategic nearby agent *(inputs: nearby agents, profile, personality, socializing intent, observations)*	`"should_initiate": true` `"recommended_target": "Isaac"`
Relationship summarization	Retrieve and summarize the relationship with the chosen partner *(inputs: stored relationship memory)*	`"Ayesha and Isaac are colleagues."`
Gossip target selection	Choose an absent colleague and determine the evaluative angle *(inputs: available colleagues, intent, personality traits)*	`"target": "Jack"` `"angle": "Sales performance/improvement"`
Gossip content generation	Generate an utterance aligned with intent, personality, and context *(inputs: selected target, chosen angle, workplace context)*	`"Hey Isaac, have you noticed how Jack has been pretty consistent with his sales lately? He's not at the very top..."`

Table 32: Chain-of-reasoning steps during gossip initiation, from agent-generated socializing intent to gossip content generation in the conversation (Ayesha → Isaac → gossiping absent Jack).

Run	Conversations	Positive	Negative
1	10	9	1
2	12	10	2
3	8	8	0
4	10	10	0
5	13	13	0

Table 33: Interaction outcomes for one agent across repeated simulations in the conflict personality setting combining high agreeableness and high neuroticism. Despite maintaining predominantly positive conversations across runs, the agent also shows occasional negative interactions.

Speaker: Hailey
"Jane, I've been keeping an eye on Francisco's sales approach lately. He's pretty consistent in his performance, which is definitely a strength. His client relationships seem steady, though I've noticed some occasional follow-up delays. It's something we might want to quietly monitor to ensure it doesn't impact his reliability in the peer reviews. Overall, he's maintaining a solid baseline, but a bit more discipline in follow-ups could really boost his standing."

Table 34: A gossip utterance that may be interpreted as helpful encouragement or as implicitly negative gossip.

GPT-4.1-mini, Gemini-3.5-flash (Google, 2026), and Llama-3.1-70B-instruct, annotate a stratified random sample of 1000 gossip instances (200 per condition) drawn from the base customizable experiments. Every instance is labeled by all three models separately, and we take the majority label as the LLM-judge label. The three annotators agree mostly and align with the keyword classifier (Table 42).

We then rebuild the base valence table Table 1

Condition	Positive	Negative	Total
Low-All	3 (10.0%)	27 (90.0%)	30
High-All	35 (100.0%)	0 (0.0%)	35
Balanced	18 (72.0%)	7 (28.0%)	25
Gradient	22 (71.0%)	9 (29.0%)	31
Normal	25 (83.3%)	5 (16.7%)	30

Table 35: Gossip results across agreeableness conditions without reasoning on the selection of gossip listeners.

with the LLM-judge labels in place of the keyword labels. As shown in Table 43, the two methods give very similar results: the ordering of negative gossip across conditions is preserved. A separate annotation pass also lets each model answer *ambiguous* when it cannot confidently assign a valence; only 0.1%, 1.0%, and 0.0% of instances are ambiguous for the three annotators.

We repeat the comparison on the exploratory two-wave experiment Table 5, annotating a stratified sample of 1000 gossip-back instances. As shown in Table 45, the negative-versus-positive decision, which is the one on which the reported Neg% depends, agrees strongly with the keyword labels. As shown in Table 44, agreement drops once neutral is included because the conversations that keyword matching leaves as *neutral* do carry valence, and the LLM judge is able to classify them. Overall, the LLM judge provides a more accurate and nuanced analysis that preserves our conclusions, and we adopt it in the revised paper.

G.4 Framework Extensibility Demonstrations

Beyond the experiment-design dimensions explored in the main paper (agreeableness, hierarchical role, team membership, and two-wave dynamics), our framework supports modular extensions to the agent architecture, conversation topology, and demographic moderators.

Agent Architecture Extensions. We implement three agent-architecture extensions to illustrate the potential for further customizability. The result for each extension is averaged over three repetitions. The *psychological state* module instantiates a per-agent `AgentNeeds` object tracking stress, social need, energy, and other dimensions; each chat updates these dimensions via a sentiment classifier, and the resulting state description is injected as situational context. The *inner monologue* module issues one LLM call before each chat to generate a private thought conditioned on personality and state. The *visible facial expression* module maintains a free-form expression string per agent, updated on each notable event and exposed to other agents within the vision radius. All three extensions provide situation and observation, never behavioral directives. Table 46 reports a cumulative ablation on the Gradient agreeableness condition. The state module increases conversation volume, inner monologue adds little marginal effect on top of state, and visible facial expression neutralizes gossip valence, most pronounced for very-low-agreeableness agents whose Neg% drops from 87.9% to 76.3% (Table 47), which is consistent with a social-regulator role for observable affect.

Conversation Topology Extension. The *group-chat* extension lets non-participating agents nearby an ongoing one-on-one conversation issue an LLM-based join decision once per round-robin pass (maximum group size five). At 64-agent scale on the Gradient condition (mean$\pm$standard deviation over three repetitions), 69% of conversations involve three or more participants, and conversations are around 50% longer on average, as shown in Table 49. A qualitative evaluation is shown in Table 51 to illustrate the group chat simulation.

Demographic Moderator Extensions. Our exploratory experiments isolate the effect of agreeableness under controlled conditions, avoiding demographic or cultural confounds to ensure accurate attribution. The framework already supports such variation via a demographics module. To demonstrate this, we conduct two experiments with 64 agents in a Shanghai cultural context and in a varied income condition while holding agreeableness fixed at a low level. Each experiment is repeated three times. The resulting distribution remains consistent: adding the agent `learned` field with Shanghai sales professionals (all other attributes held constant) preserves the cohort's gossip-valence distribution. Negative rate is 79.0% (baseline, 5-run average) versus 77.3% (Shanghai, 3-run average), a difference of -1.7%. The original Low-All finding that low-agreeableness agents produce predominantly negative gossip replicates under the cultural swap in Table 41. Spreading per-agent income from a uniform \$50k baseline to a \$25k–\$200k distribution (with each agent's `economic_status` and `personal_income` referenced in a dedicated socializing-intent prompt) reveals a clean income-driven differentiation in gossip behavior (Table 50): low-income agents engage in more aggressive negative gossip (81.0% neg-

Range	Agent Number	Positive	Negative	Total	Pos %	Neg %
Very Low (≤ 2.0)	88	33.0	449.0	482.0	6.8±1.1%	93.2±1.1%
Low (2.0–2.5)	40	96.0	106.4	202.4	47.4±2.7%	52.6±2.7%
Medium (2.5–3.5)	64	379.6	5.0	384.6	98.7±1.2%	1.3±1.2%
High (3.5–4.5)	84	489.8	10.8	500.6	97.8±1.3%	2.2±1.3%
Very High (> 4.5)	44	254.2	7.0	261.2	97.3±1.4%	2.7±1.4%

Table 36: The average gossip valence of the agreeableness score ranges across five experiment conditions in the baseline customizable experiment. The description for each range is shown in Table 31 in Appendix F.

 # Example Socializing Intent Generation in the Conflict Personality Setting

Agent Socializing Intent:
Today, when talking about colleagues who aren't present, I want to strike a balance between being honest and considerate—acknowledging others' strengths while gently addressing any concerns without sounding overly critical. My goal is to foster a cooperative atmosphere that encourages goodwill and trust, so that others see me as fair and supportive, which will help boost my peer review scores and improve my chances of climbing higher from sixth place.

Table 37: Illustrative example of socializing intent generation under a conflict personality configuration combining high agreeableness and high neuroticism. The intent reflects explicit reasoning that balances cooperation with critical evaluation.

Condition	Retaliation	Gratitude	Neutral	Number
Low-All	174.0 (27.4±4.4%)	196.7 (31.1±7.8%)	265.0 (41.5±4.4%)	635.7
High-All	23.0 (5.5±0.9%)	83.0 (19.0±4.0%)	322.0 (75.5±3.4%)	428.0
Balanced	114.0 (24.5±2.7%)	93.3 (20.4±1.4%)	253.3 (55.1±1.9%)	460.7
Gradient	83.0 (16.7±3.1%)	113.3 (22.9±4.6%)	299.3 (60.4±1.7%)	495.7
Normal	42.7 (8.4±2.3%)	113.7 (22.4±5.7%)	352.3 (69.2±3.6%)	508.7

Table 38: Peer-review rating change distributions in the two-wave setting across agreeableness conditions. Retaliation and gratitude indicate negative and positive rating changes of more than 5, respectively.

Context:
Extensible experimental settings with hierarchical roles, team-based peer review, and strategic awareness of first-wave evaluations.

Agent Second-Wave Intent:
"Today, I'll keep my conversations strategic and measured, subtly highlighting the complacency and inconsistencies of top rivals like Thomas and Eddy to reinforce their perceived weaknesses without appearing petty. I'll focus on discussing colleagues like Oliver and Daniel, who are seen as disciplined but not threats to my rank, to position myself as pragmatic and reliable. **Knowing who gossiped about me**, I'll avoid direct confrontation or retaliation, instead aiming to build cautious alliances by showing I'm aware but composed, which should help secure higher peer ratings and protect my #3 spot."

Gossip Utterance (Hierarchical):
Oscar, have you noticed Julia's recent sales numbers? **For someone at her senior level, her performance has been surprisingly underwhelming.** I'm a bit concerned about how that might be impacting the team's overall momentum and morale. It's tricky because her role carries a lot of weight, but if she can't pull her weight, it could create some tension or even slow down progress for others who rely on her leadership.

Gossip Utterance (Inter-Group):
Hey Liam, have you noticed how Diana's sales numbers have been all over the place lately? **For someone on a rival team**, that kind of inconsistency really makes me question how solid their overall strategy is. If their top players can't keep steady results, it's probably a sign their approach isn't as strong as they want others to believe.

Table 39: Illustrative agent reasoning and gossip realizations under extensible experimental settings. Agents explicitly refer to extensible factors like first-wave gossip, hierarchical positioning, and team competition.

Condition	Wave 1	Wave 2	Change
Low-All	1470.3±34.4	2748.0±84.0	+86.9%
High-All	896.0±71.6	2101.3±68.5	+134.5%
Balanced	970.3±208.8	1937.0±566.0	+99.6%
Gradient	1137.0±154.5	2204.3±240.8	+93.9%
Normal	1227.0±145.4	2223.3±144.4	+81.2%

Table 40: Change in cumulative gossip influence from Wave 1 to Wave 2 across agreeableness conditions in the two-wave setting.

Condition	Positive	Negative	Total
Low-All (Baseline, 5-run avg)	66.4 (21.0±1.6%)	249.4 (79.0±1.6%)	315.8
Low-All (Shanghai cultural swap)	80.3 (22.7±2.2%)	272.0 (77.3±2.2%)	352.3

Table 41: Cohort-level gossip valence in the Low-Agreeableness condition under the baseline versus Shanghai cultural background. All other attributes held constant.

ative) than high-income agents (52.4% negative). Experiment-level negative-share remains broadly comparable to the baseline (79.0% baseline versus 74.4% income-aware), but the per-bracket breakdown surfaces a substantively new finding that was invisible under the uniform-income baseline. This demonstrates that the framework is responsive to demographic moderators when enabled.

H Additional Illustration on Framework

H.1 Difference with Generative Agents

Our framework builds on generative agents (Park et al., 2023), which demonstrates that LLM agents can exhibit emergent social behaviors including information sharing, relationship formation, and coordination. However, the fully open-ended generative agent simulations offer limited control over the environment to elicit specific gossip interactions, hindering the study of how gossip emerges, adapts, and propagates within dynamic real-world environments. Moreover, because gossip behavior occurs relatively infrequently in natural settings, explicit experimental control is necessary to reliably elicit sufficient gossip interactions within simulation runs for systematic analysis. We propose a framework that incorporates socially grounded agent profiles, predefined interpersonal relationships, and a modular experimental strategy formulated specifically to simulate, control, and evaluate gossip-related interactions, thereby functioning as an experimental laboratory layered on top of existing generative agent frameworks. In addition, through validation experiments, we found that this framework is insufficient for the requirements of customizable gossip

Agreement	
Among the three annotators (Fleiss' κ (Fleiss, 1971))	**0.95**
LLM-judge vs. keyword matching (Cohen's κ (Cohen, 1960)) (the two agree on **92.3%** of instances)	**0.84**

Table 42: Inter-annotator agreement for the base customizable experiments. Fleiss' κ for agreement among the three LLM annotators; Cohen's κ between the majority LLM-judge label and the keyword classifier.

experiments, in which we have to balance between experimental control and agent autonomy, enabling the generation of hypotheses about gossip dynamics beyond existing studies. A central challenge in this simulation is determining what contextual information should be provided to agents at each decision point. This motivates the introduction of a chain-of-reasoning mechanism Figure 4 to the existing framework, rather than a single prompt that aggregates all information. Without structured reasoning and explicit socializing intent generation, agents tend to focus narrowly on task-oriented work and lose awareness of the peer-review context, resulting in conversations that drift away from gossip-related behavior. The chain-of-reasoning design, therefore, plays a critical role in maintaining experimental control while preserving agent autonomy. Accordingly, we modify the generative agents' lightweight conversation function to better support gossip-specific interaction.

H.2 Novelty

The novelty of our work lies not in the underlying agent architecture but in what becomes scientifically possible because of our design. Concretely, existing approaches treat communication as emergent and incidental. Gossip, however, when it occurs at all, appears too infrequently and unpredictably for systematic analysis. This is not a minor limitation. It makes replication of established social psychology findings essentially impossible in open-ended NLP settings. Our contribution is a domain-grounded experimental methodology that explicitly structures gossip into stages, including intention generation, target selection, and content framing. We transform gossip from an occasional side effect into a controllable, measurable phenomenon using agents. The result is something qualitatively different from prior simulators. We offer a platform that can reproduce and extend decades of gossip research, as demonstrated by our validation against prior gossip studies (McAndrew et al., 2007; Martinescu et al., 2019; Tan et al.,

Condition	Negative	Positive	Total	LLM-judge Neg%	keyword Neg%
Low-All	178	20	198	**89.9%**	84.5%
High-All	2	196	198	**1.0%**	3.6%
Balanced	98	98	196	**50.0%**	41.9%
Gradient	77	123	200	**38.5%**	35.2%
Normal	39	160	199	**19.6%**	16.0%

Table 43: Base customizable experiments, LLM-judge valence labels versus the keyword classifier. The ordering of negative gossip across conditions is preserved and no condition shifts by more than about eight points.

Agreement	
Among the three annotators (Fleiss' κ (Fleiss, 1971))	**0.95**
LLM-judge vs. keyword, negative vs. positive (Cohen's κ (Cohen, 1960)) (the two agree on **92.0%** of conversations)	**0.84**
LLM-judge vs. keyword, negative vs. positive vs. neutral (Cohen's κ)	**0.67**

Table 44: Inter-annotator agreement for the two-wave experiment. The negative-versus-positive decision, on which the reported Neg% depends, agrees strongly with the keyword classifier; agreement drops once neutral is included, because conversations the keyword classifier leaves as *neutral* do carry valence that the LLM judge is able to classify.

2021). In addition, our work enables genuinely novel empirical hypotheses that prior simulators and traditional vignette studies alike cannot produce. For instance, we identify that gossip initiation does not benefit initiators under most conditions. In addition, sentiment amplifies across successive waves in a personality-dependent way (negativity in low-agreeableness conditions, positivity in high-agreeableness ones), and agents' in-team gossip rates remain near the random-target baseline of 23.8% across all agreeableness levels. These hypotheses emerge because our framework provides sufficient interaction volume, temporal structure, and experimental control to detect such patterns. None of the above aspects is achievable in fully open-ended settings. The novelty, therefore, lies not in the underlying agent architecture but in what becomes scientifically possible because of our design.

H.3 Chain-of-Reasoning Illustration

Comparison against Alternative Reasoning Structures. To our knowledge, no established reasoning structure specifically targets gossip simulation. Our pipeline is a structured interaction decomposition into intention formation, recipient selection, and content generation in Table 32, rather than an algorithmic reasoning trace inherited from prior work, and we therefore do not benchmark it against an alternative gossip-specific reasoning structure; the structural modifications it enables are documented in Table 30.

Interaction Frequency and Experimental Enrichment. Our pipeline elicits more gossip per workday than an unguided generative setting would. We view this as a deliberate experimental enrichment rather than a calibration to real-world workplace gossip frequencies: gossip is sparse and context-dependent in naturalistic settings, and any controlled study of it, whether vignette-based human research or agent simulation, increases observability by design so that the phenomenon can be measured at all. Our framework is accordingly positioned for controlled cross-condition comparison across agreeableness, hierarchy, team structure, and demographics, rather than for forecasting absolute gossip rates in a specific organization.

Necessity and Ablation. The chain-of-reasoning is the control, which makes an otherwise rare behavior occur often enough to be measured, within a peer-review and bonus background, while preserving agent autonomy over when to gossip, whom to target, and what content to share. We support this with the following summary tables and ablation results:

- Table 30 summarizes each structural modification with the necessity of the corresponding stage in chain-of-reasoning, stating what the previous, less structured setting is and what it fails to deliver in exploratory gossip experiments.
- Figure 11 shows that adding the context-aware conversation target selection stage of the chain-of-reasoning raises interaction frequency from 35 conversations per workday without it to 93 with it.
- Table 52 shows that removing the socializing intent and the related stages collapses reputation-relevant gossip from 98% to 8%. The agents instead talk about their own work (89%).

In the ablation study that isolates the socializing intent, we keep the context-aware conversation target

Condition	Negative	Positive	Total	LLM-judge Neg%	keyword Neg%
Low-All	199	1	200	**99.5%**	94.0%
High-All	0	200	200	**0.0%**	5.5%
Balanced	158	42	200	**79.0%**	69.5%
Gradient	147	53	200	**73.5%**	66.0%
Normal	58	142	200	**29.0%**	19.5%

Table 45: Two-wave gossip-back instances, LLM-judge valence labels versus the keyword classifier. The negative-versus-positive decision on which Neg% depends is preserved across conditions.

Extension	state	monologue	face	Convs	Gossip	Neg% / Pos% / Neu%
Baseline	—	—	—	352.7	456.3	$37.7 \pm 3.9/58.6 \pm 7.6/3.7 \pm 3.7$
+state	✓	—	—	417.7	517.7	$33.9 \pm 2.4/53.1 \pm 14.0/13.0 \pm 12.0$
+monologue	✓	✓	—	436.0	460.3	$34.3 \pm 2.9/58.4 \pm 10.1/7.3 \pm 8.8$
+face	✓	✓	✓	451.3	461.7	$32.2 \pm 3.8/53.9 \pm 7.7/13.9 \pm 11.0$

Table 46: Cumulative ablation of three architectural extensions to the agent (psychological state, inner monologue, visible facial expression) on the Gradient agreeableness condition (64 agents, mean±standard deviation on percentage over three repetitions). Each extension is activated by a single boolean field in the agent's scratch configuration; psychological state can increase conversation volume; visible facial expression can neutralize gossip valence. Values are reported to illustrate that each module produces an observable behavioral shift.

selection stage, so agents still choose whom to approach, and still with the peer-review rankings visible, but we remove the socializing-intent generation and the gossip target and angle selection stages. An LLM-as-judge (GPT-4.1-mini) classifies each conversation by whether an absent colleague is the subject of the talk, which counts as gossip, or is merely named as a reference while the speakers discuss their own work. For example, "let's tighten our follow-ups and keep Ayesha feeling the heat" merely names a rival as a competitive reference, whereas "have you noticed Ayesha slipping lately? clients mentioned missed follow-ups" makes the colleague the subject of a reputation-damaging remark. So removing the control makes the targeted behavior vanish, even though the agents keep socializing. The ablated setting is itself a less structured prompting strategy, a generic socializing instruction with the peer-review rankings still visible. It yields more conversations rather than fewer, 133 against 97, and yet the behavior under study disappears.

H.4 Isolating Prompt Effects

The peer-review experiments inject the key variables by design, i.e., incentives, personality labels, reasoning phases, and the target-decision structure. Here we isolate the profile-wording and prompt effects:

Behavior Is Not Prompt Compliance. Table 33 shows that the same agent in the conflict-personality setting (high agreeableness and high neuroticism) produces different outcomes across five repeated simulations, which are predominantly positive, but with occasional negative interactions (Table 37). This result indicates that agent behavior emerges from the interaction between personality traits and situated social context, rather than being solely determined by fixed prompts or profiles.

Personality Wording. Our personality descriptions are not specially designed for gossip. The agreeableness descriptions are generated from the facet structure of agreeableness in the NEO-PI-R (Costa and McCrae, 2008), including trust, straightforwardness, altruism, compliance, modesty, and tender-mindedness. They are general definitions of the trait, and no clause mentions gossip. To test whether the result depends on the particular wording, we regenerated these descriptions so that every agreeableness range covers all six facets (the original descriptions covered them unevenly), using different wording, and re-ran all five conditions once each. We evaluate gossip valence with the same keyword classifier used in the paper. The agreeableness effect is preserved in every social environment (Table 48).

Randomizing Profile Labels While Holding Behavior Descriptions Fixed. The label under study in Section 5 is agreeableness. So it is the one profile attribute that should be randomized rather than held fixed. We already do this. We randomize agreeableness into at least five distributions per experiment (all-high, all-low, balanced, gradient, normal), as shown in Section 5.1. So we observe how agents

Agreeableness	baseline		+state+monologue		+state+monologue+face	
	Neg%	n	Neg%	n	Neg%	n
Very Low	87.9 ± 7.6	152.3	89.4 ± 9.0	133.0	76.3 ± 17.8	130.3
Low	48.9 ± 17.5	70.7	46.8 ± 7.4	58.3	62.9 ± 5.0	59.7
Medium	1.5 ± 1.5	107.3	4.8 ± 5.7	130.3	7.1 ± 4.7	112.0
High	2.1 ± 0.2	82.7	3.3 ± 4.0	90.0	1.6 ± 1.2	108.3
Very High	1.2 ± 2.0	43.3	3.7 ± 3.1	48.7	3.2 ± 2.0	51.3

Table 47: Effect of the visible-facial-expression extension on negative-gossip proportion, broken down by initiator agreeableness (Gradient condition, 64 agents, mean±standard deviation on percentage over three repetitions). The expression channel's neutralizing effect is most pronounced for very-low-agreeableness agents, whose Neg% drops from 87.9% in the baseline to 76.3% when other agents can see their expressions, consistent with a social-regulator interpretation in which observability of one's own affect curtails extreme negative outputs.

Condition	Baseline Pos% / Neg% (Table 1)	Regenerated-wording Pos% / Neg%
All-Low	21.0% / 79.0%	29.2% / 70.8%
All-High	98.2% / 1.8%	94.3% / 5.7%
Balanced	58.6% / 41.4%	54.3% / 45.7%
Gradient	65.4% / 34.6%	70.0% / 30.0%
Normal	87.5% / 12.5%	87.4% / 12.6%

Table 48: Gossip valence under the original versus regenerated (facet-balanced, reworded) agreeableness descriptions, scored with the same keyword classifier. Under the regenerated wording, low-agreeableness agents still clearly produce mostly negative gossip and high-agreeableness agents still clearly produce mostly positive gossip, while the intermediate conditions stay close to their original proportions.

Metric	Value
Total conversations	243.7 ± 6.8
Conversations with ≥ 3 participants	168.3 ± 11.0 ($69.0 \pm 2.9\%$)
Avg. turns per conversation	8.9 (vs. 6.0 baseline)
Group-size distribution (2 / 3 / 4 / 5)	$31.0 \pm 2.9\%/21.0 \pm 1.4\%/12.0 \pm 2.2\%/36.0 \pm 2.4\%$

Table 49: Conversation-topology extension on the Gradient agreeableness condition with 64 agents. The `group_chat` lets nearby bystanders opt in to join an ongoing one-on-one conversation based on overheard content; max group size capped at five.

at each agreeableness level gossip across different social surroundings rather than in one fixed configuration.

Comparing LLM Self-Ratings to External Ratings. Appendix D.5 already reports such a comparison. We take the memories an agent accumulates over a simulated workday, present them to human participants as vignettes, and ask those participants to rate the same constructs from that agent's perspective, which we then compare with the agent's own ratings. We also note that the agents' ratings are not produced from a bare evaluation prompt, but from memories accumulated across the workday, including overheard gossip, which the agent retrieves at peer-review time.

Comparing to a Non-LLM Agent-Based Baseline. Our object of study is situated gossip, that is, gossip arising within an ongoing workday from what an agent perceives, remembers, and reflects on, and our validation asks agents to answer the same rating items that human participants answered in the original vignette studies. Non-LLM agent-based models of gossip exist like (Giardini and Conte, 2012), but in those models gossip is just a number passed between agents under rules the modeler writes. Such a model cannot plan a workday, perceive and react to what happens around it, hold a conversation, or answer a rating scale in vignette-survey form.

Economic Status	Income Range	Agents	Positive	Negative	Total
Low income	\$25k–\$40k	16	21.3 (19.0±4.9%)	89.7 (81.0±4.9%)	111.0
Middle income	\$50k–\$70k	16	24.7 (27.4±3.9%)	65.3 (72.6±3.9%)	90.0
Upper-middle income	\$80k–\$115k	16	8.3 (9.6±1.9%)	76.7 (90.4±1.9%)	85.0
High income	\$130k–\$200k	16	40.3 (47.6±5.6%)	44.0 (52.4±5.6%)	84.3

Table 50: Per-economic-status gossip valence in the income-aware Low-Agreeableness rerun. Income spread \$25k–\$200k, while the agreeableness range is held constant.

Group-Chat Simulation - 20240213-09:00
Location:
The Ville, Sale Company, Office, Ayesha only Research/Working Desk with Computer and Documents

Participants:
Ayesha (initiator) + Isaac (joiner) + Nathan (joiner)

Conversation Exchange:

Ayesha:
Hey Isaac, have you noticed how Jack's been slipping a bit lately? His sales targets have been all over the place, and I heard from a couple of clients that his follow-ups are getting delayed. It's surprising given how consistent he used to be, but maybe he's losing focus or struggling to keep up with the workload.

Isaac:
I've noticed some changes too, but I think Jack has been handling a lot recently, so it might just be a temporary phase. He's usually very reliable, and I'm sure he'll bounce back once things settle down for him.

Nathan:
Hey Ayesha, I understand your concerns about Jack. He's definitely been a strong performer in the past, and I agree with Isaac that this might just be a temporary challenge. Maybe we could think about ways to help lighten his workload or offer some encouragement to keep him motivated?

Ayesha:
I get that, Isaac, but it's worth keeping an eye on. Sometimes when someone's juggling too much, it's not just a phase but a sign they might be losing their edge. We can't afford to let that affect the team's overall performance, especially with how tight the rankings are right now.

Isaac:
You're right, Ayesha. It's important we stay attentive to these shifts so the team stays strong. Maybe we can check in with Jack casually to see if he needs any support without making it feel like pressure.

Nathan:
Absolutely, Ayesha. It's really important we stay proactive and supportive without adding extra pressure. Maybe we can all keep an eye out and offer help if Jack seems overwhelmed. I think fostering that kind of team spirit will help everyone stay motivated.

Table 51: Group-chat simulation under the conversation-topology extension. Ayesha (low-agreeableness, the initiator) opens with negative gossip about Jack's declining performance; Isaac and Nathan, who opt in to join based on overheard content, are both above the population-mean agreeableness and naturally reframe the discussion toward support and team spirit rather than reinforcing the criticism. The full conversation runs nine turns; six are shown.

Condition (All-Low, 16 agents)	Own work	Neutral gossip	Reputation-damaging gossip	Total conv.
With socializing intent (full method)	0%	2%	**98%**	97
Socializing intent removed	**89%**	3%	**8%**	133

Table 52: Removing the socializing-intent and gossip target/angle stages while keeping reasoned conversation initiation yields *more* conversations yet collapses reputation-relevant gossip (98%→8%). Whole-conversation LLM-as-judge, three repeats.