Multi-Agent Comedy Club: Investigating Community Discussion Effects on LLM Humor Generation

Anonymous ACL submission

Abstract

Prior work has explored multi-turn interaction and feedback for LLM writing, but evaluations still largely center on prompts and localized feedback, leaving persistent *public* reception in online communities underexamined. We investigate whether broadcast community discussion improves stand-up comedy writing in a controlled multi-agent sandbox: in the discussion condition, critic/audience threads are recorded, filtered, stored as social memory, and later retrieved to condition subsequent generations, whereas the baseline omits discussion. Across 50 rounds (250 paired monologues) judged by five expert annotators using A/B preference and a 15-item rubric, discussion wins 75.6% of instances and improves Craft/Clarity (Δ=0.440) and Social Response (Δ=0.422), with occasional increases in aggressive humor.

1 Introduction

Large Language Models (LLMs) are increasingly deployed as writing assistants that personalize outputs using authors' historical documents and generate actionable feedback for revision (Mysore et al., 2024; Chamoun et al., 2024). Given that LLM-generated texts are now prevalent on social media and significantly impact human readers (Radivojevic et al., 2024), a natural hypothesis arises: networked discussions may, in turn, influence LLMs. This explicitly mirrors core dynamics in online communities, where creative writing is inherently linked to public reception (e.g., comments and critiques) that authors use to refine their work (Guo et al., 2023; Cheng and Frens, 2022).

Motivated by this perspective, we ask a concrete question: *can broadcast community discussion be operationalized as a usable conditioning signal that improves an LLM's subsequent creative writing?* To answer this, we build a controlled multi-agent sandbox that instantiates a small stand-up comedy community and allows us to manipulate whether public reception is generated, logged, and fed back into later rounds (Figure 1). Humor provides a demanding testbed for reception-grounded generation since stand-up comedy is explicitly audience-oriented, and success is defined by audience reaction (Mirowski et al., 2025).

Meanwhile, agentic improvement paradigms wrap an LLM in iterative generate–evaluate–revise loops and typically rely on *private, self-generated* feedback (e.g., Reflexion) (Shinn et al., 2023). In contrast, we study a controlled setting in which reception is (i) *public and broadcast*, (ii) *logged as an interaction trace*, and (iii) *reused across rounds through a bounded interface* (e.g., a fixed-size retrieved memory window). By holding within-round generation constant and varying only the cross-round reception stream, we can attribute improvements to accumulated public feedback rather than to within-round editing or private self-critique common in agentic loops.

We build such a setting inspired by the "Smallville" sandbox, where LLM agents sustaining routines, social interaction, and memory in a small community (Park et al., 2023). We instantiate a stand-up comedy community: a host releases a fixed sequence of prompts, five performer agents produce stand-up monologues, and a community of critics and audience agents responds through threaded discussion. Our intervention is a binary switch that either enables or skips the post-performance discussion phase. Performers write exactly once per round and do not revise within the round, so any effect of reception can only manifest *across rounds* via logged discussion and a bounded social-memory interface that retrieves relevant reception into later contexts. This bounded interface is also motivated by known limitations in how LLMs use long contexts (Liu et al., 2024).

To evaluate community discussion effects on the outcome, we conduct a dedicated *human evaluation* of paired outputs from the two conditions,

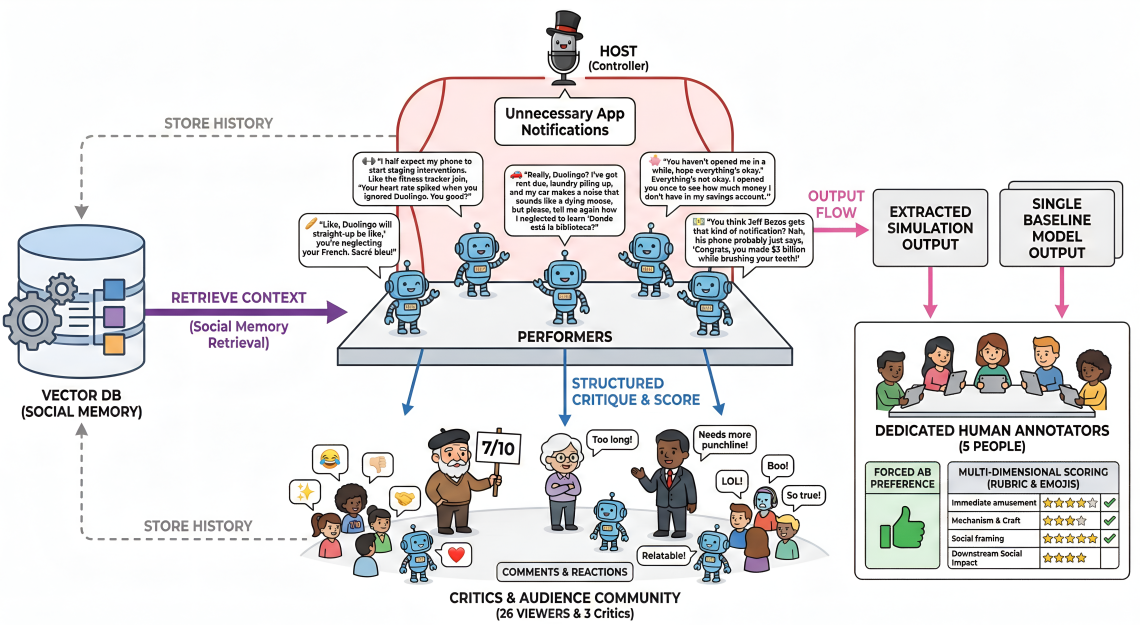

Figure 1: Overview of **Multi-Agent Comedy Club**. In each round, a host prompts five performer agents to write stand-up comedy monologues. When enabled, a broadcast discussion produces threaded reception (critique, scores, and reactions) that is stored in social memory and retrieved to condition later rounds. We extract paired outputs from the simulation and a baseline model without discussion simulation, and evaluate them with dedicated human annotators via forced A/B preference and multi-dimensional rubric ratings.

using forced-choice A/B preference and multi-dimensional rubric ratings spanning (i) outcomes (preference and immediate amusement), (ii) mechanism & craft, and (iii) social reception. Across 50 rounds (250 paired monologues), the discussion-enabled condition wins 75.6% of instances and yields consistent gains in Craft/Clarity (Δ=0.440) and Social Response (Δ=0.422), indicating that reception-grounded conditioning can improve long-form creative writing even without within-round revision. At the same time, improvements can come with stylistic tradeoffs (e.g., shifts toward edgier humor), motivating a multi-objective view of quality versus social risk. We will release our sandbox configuration, paired outputs from both conditions, and reconstructed discussion threads to support replication and future work.

Contributions. This paper makes three contributions: (1) **Sandbox Mechanism**, a controlled paradigm for *reception-grounded* creative generation; (2) a **paired resource** of long-form stand-up monologues under matched conditions and community discussion threads; and (3) a **diagnostic human evaluation protocol** for long-form humor.

2 Related Work

We review prior work on (i) computational humor evaluation and generation, (ii) multi-agent interaction and feedback for creative writing, and (iii) agent-based social simulation. Together, these studies motivate our focus on public reception as an explicit interaction signal that can cumulatively shape long-form humor writing across rounds.

2.1 Computational Humor

Humor serves as a critical metric for evaluating creativity and contextual processing in artificial intelligence. Classical semantic theories define humor as the simultaneous presence of conflicting scripts within a single text, where the humorous effect arises from a cognitive shift or reinterpretation between these frameworks (Raskin, 1979). Consequently, humor comprehension and generation tasks have become effective benchmarks for assessing LLM capabilities (Loakman et al., 2025). Since humor relies on implicit cultural knowledge and nuanced associations, recent studies utilize humor comprehension to evaluate reasoning abilities that extend beyond conventional STEM benchmarks (Narad et al., 2025; Cocchieri et al., 2025).

Empirical evidence indicates that LLMs can achieve competitive performance in generating short-form jokes under constrained prompts (Gorenz and Schwarz, 2024; Cao et al., 2025). However, isolated prompting struggles to steer or evaluate the voice, pacing, and narrative payoffs required for long-form comedy. Accordingly, we complement humor judgments with creative-

writing and social rubrics to diagnose both craft and social impact in the current research.

2.2 Multi-agent Interaction and Feedback for Creative Writing

A growing body of literature conceptualizes creative writing as a collaborative process involving multiple agents, exemplified by frameworks simulating writers' rooms (Huot et al., 2025), director-actor dynamics (Han et al., 2024), and character-driven storytelling (Yu et al., 2025; Ran et al., 2025). Role-playing benchmarks further emphasize the importance of controllable personas and distinct speaking styles in stabilizing role-consistent behavior and fostering diverse perspectives (Wang et al., 2024a). In the specific domain of humor, feedback serves as a supervision signal beyond mere imitation; Ravi et al. (2024) demonstrate that assigning dual roles of teacher and critic helps narrow the distillation gap in humor generation.

Despite these advancements, multi-agent interaction does not yield uniform benefits. In certain reasoning contexts, a well-prompted single agent can match or even exceed multi-agent performance (Wang et al., 2024b). These mixed findings suggest that "multi-agent" is not a mechanism by itself; the mechanism is the feedback channel that becomes available and reusable. Consequently, we isolate reception as the primary intervention in our study.

2.3 Agent-based Social Simulation

LLM agents are increasingly studied in simulated environments that produce multi-turn interaction traces for analyzing behavior and collective dynamics. Foundational work such as Park et al. (2023) demonstrates how memory and reflection can yield emergent routines, while newer simulators adopt platform-mimetic structures to study social influence and recommender-mediated phenomena (Puelma Touzel et al., 2025; Wang et al., 2025) and to explore population-level interventions (Piao et al., 2025; Mi et al., 2025). Recent surveys and benchmarks further emphasize fidelity, reasoning structure, consistency, safety, and coordination effects as core evaluation concerns (Gao et al., 2024; Li et al., 2025; Zhu et al., 2025).

However, prior social simulators seldom use controlled, paired designs to test how a manipulable interaction channel causally shapes *creative* outputs. We address this gap by directly manipulating whether reception within community is generated and fed back into later contexts and making public discussion an explicit experimental factor for multi-round writing improvement.

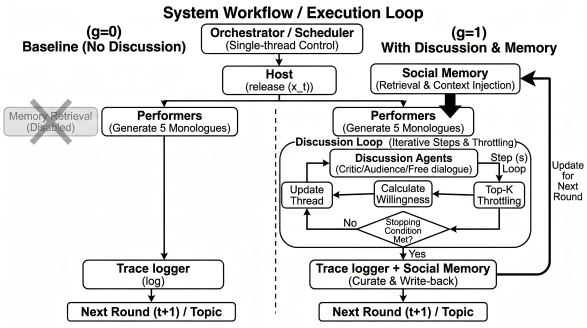

Figure 2: Workflow overview of our multi-agent sandbox. Left: baseline (g=0) skips discussion and logs performances only. Right: community discussion (g=1) adds an iterative discussion loop that produces reception, which is written to social memory at the end of round t and retrieved to condition performers at the start of round t+1.

3 Sandbox Simulation: Multi-Agent Comedy Club

We design a closed comedy community sandbox to study reception-grounded writing under experimental control over the topic list, model, and agent identities constant, so that observed differences are plausibly attributable to community discussion. This section details the experimental manipulation and provides a workflow overview (Figure 2).

3.1 Settings and Variables

The system runs in discrete rounds indexed by t. In each round, a host releases a topic prompt x_t and five performer agents each produce one monologue of stand-up comedy.

Our manipulated factor is whether performances are followed by a *broadcast community discussion* ($g = 1$) or not ($g = 0$). In the community discussion condition ($g = 1$), performances are followed by a discussion phase that produces critic reviews, audience posts, and free dialogue organized as threaded discussions. In the baseline condition ($g = 0$), we instantiate the same agent roster and roles, but we skip all non-performer stages. After logging the five performances for topic x_t, the system directly advances to topic x_{t+1}.

Performers do not revise within a round, so any effect of community reception can only occur across rounds via logging reception, writing it into social memory, and retrieving it into later performer contexts.

3.2 Agents, Personas, and Model

The sandbox contains N=35 agents with fixed persona text: five performers, three critics, twenty-six audience members, and one host. All agents are instantiated using the same GPT-4o-mini model. Across conditions, we keep the decoding configuration fixed and use the same role-specific output length caps. Input contexts differ by design because the community discussion condition provides additional observable discussion content.

Personas. Persona text specifies role and voice. We use personas to (i) stabilize role-consistent behavior across rounds, making reception signals easier to interpret, and (ii) encourage diverse viewpoints in discussion without adding extra control rules. Full persona text is provided in Appendix C.

3.3 Round Protocol and Discussion Dynamics

Topic control. We pre-generate a fixed topic list $\{x_1, \ldots, x_{50}\}$ once and reuse the same list in both conditions. In round t, the host releases the same topic x_t to the performers in both conditions.

Phase 1: Topic release. The host publishes x_t.

Phase 2: Performances. The five performers generate monologues in a fixed order from 1 to 5. Each performer generates exactly one monologue and there is no within-round revision.

Phase 3: Community discussion (g=1 only). Critics produce official reviews and audience agents produce posts. Agents may continue free dialogue in the same threaded space until a stopping rule ends the round. A *thread* is the unit of community reception. Figure 3 illustrates what constitutes a thread in our setting. Event logging and thread reconstruction are specified in Appendix A.

Step definition and willingness. The free dialogue phase is agent-driven. At each dialogue step s, every agent except the host receives its persona text, the round topic, and bounded context $C(a, t, s)$ in Sec. 3.4 and outputs JSON including a willingness score $w(a, t, s) \in [0, 1]$ and an optional `replyTo` (agent name). If w is low, the agent may output empty content.

Adaptive throttling. To keep discussion readable, we enforce adaptive throttling with $K_{\max}$=5. After collecting willingness scores, we admit

$$K_{t,s} = \min\Big(K_{\max},\ \big|\{a : w(a, t, s) \geq 0.7\}\big|\Big)$$

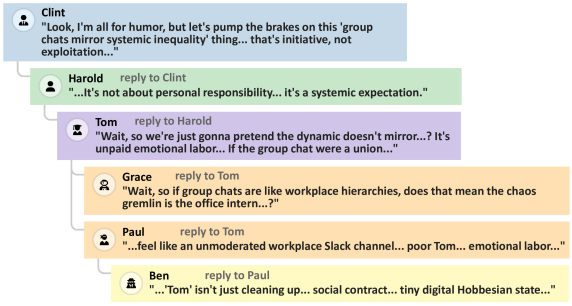

Figure 3: Visualization of a discussion thread in our setting. A thread groups reception events that are topically and referentially linked, including an initiating post (e.g., a critic review) and subsequent audience posts or free-dialogue replies.

agents: the top $K_{t,s}$ by willingness among those with $w(a, t, s) \geq 0.7$ (ties broken deterministically). Selected agents post messages; remaining agents stay silent for that step.

Stopping rule. Free dialogue terminates when either (i) the round reaches 150 free-dialogue events, or (ii) there are 15 consecutive silent steps (i.e., $K_{t,s} = 0$).

3.4 Bounded Context and Social Memory

Bounded context builder. For any agent a at round t (and dialogue step s when applicable), we build a bounded context $C(a, t, s)$ by concatenating: (i) *role anchors* (the current topic x_t, and when relevant the target performance or target thread being reacted to), (ii) a *short-term buffer* (the last L=10 utterances in the relevant thread, or the last L global utterances if no thread is specified), and (iii) an optional *retrieved community memory* block. We truncate to a fixed total budget, allocating B_{mem}=1600 tokens to the retrieved memory block and truncating at sentence boundaries. The builder structure and token budgets are identical across conditions.

Memory write-back. We implement community memory as a vector database to support cross-round conditioning. After each broadcast round (g=1), the system curates and writes high-signal reception items into the vector store. Concretely, we iterate over reception events in the raw trace (critic reviews, audience posts, and free-dialogue turns), select high-signal items (e.g., explicit critique/advice, recurring praise or complaints, and concise thread summaries used for storage), and store each item as text m with an embedding vector v, meta-

data (type, round index, and target performer when applicable), and an importance scalar π. In the baseline condition (g=0), discussion is skipped, so no reception artifacts are produced and the memory store remains empty.

Memory retrieval and ranking. When constructing $C(a, t, s)$, we retrieve community memory via embedding-based similarity search. We form a query string by concatenating the current topic, the agent persona text, and role anchors, and compute a query embedding $\mathbf{q}$ and rank memory items by

$$\text{Score}(m) = \lambda \cos(\mathbf{q}, \mathbf{v}) + (1 - \lambda)\pi + \gamma \,\text{Recency}$$

following the general design of importance and recency based retrieval (Park et al., 2023). We retrieve the k=30 highest-scoring items, then pack retrieved items into the memory block under the fixed budget B_{mem} and inject this block into $C(a, t, s)$.

Across-round conditioning for performers. Before performer P_i generates a monologue for topic x_t, we build $C(P_i, t, \cdot)$ with the current topic and retrieved community memory. This yields the intended causal chain: broadcast reception leads to the curated memory, which is retrieved and conditions performer agents' next-round writing. We omit self-reflection/revision stages to keep reception-grounded retrieval as the only cross-round mechanism. Adding explicit reflection would introduce an additional intervention and extra model calls, confounding whether gains come from community feedback versus self-critique.

3.5 Data Collection

We first run the sandbox with community discussion for 50 rounds and extract all performer monologues from the trace, yielding 250 monologues in total (5 performances per round). Across these runs, the event log contains 5,384 interaction events (including topic releases, performances, critic reviews, and free dialogue). We then run the baseline condition on the same 50 topics with the same performer roster and fixed decoding configuration, yielding a paired set of 250 baseline monologues. Each monologue is long-form, averaging ∼1,200 words.

4 Evaluation

Human evaluation is the gold standard due for humor *generation* due to its subjectivity (Amin and Burghardt, 2020). We evaluate whether multi-agent community discussion improves stand-up comedy writing over rounds with dedicated human raters. As the texts are LLM-generated, we avoid LLM-based judges to reduce correlated errors and self-evaluation bias.

4.1 Human Evaluation Metrics

Design Goal & Procedure. We frame the task as **creative writing in a social setting**: the output is simultaneously a piece of comedic writing and a social act shaped by community reception. Accordingly, we design a diagnostic evaluation to measure the impact of **community discussion** along three axes—(i) *outcomes* (does it amuse), (ii) *mechanism & craft* (how the writing lands and is constructed), and (iii) *social reception* (how it positions the speaker and propagates). For each prompt, participants select a preferred text (A/B) (**Q0**) and then rate each text on 1–5 Likert-type items (1 = strongly disagree / not at all; 5 = strongly agree / very much) (Likert, 1932).

Outcome & Mechanism/Craft Profile. We measure **Immediate Amusement** (Q1) as the primary success metric. To diagnose *how* discussion changes writing beyond raw amusement, we assess: **Reframing/Insight** (Q2), **Intent Clarity** (Q3), **Justified Landing** (Q4; coherence/justification), **Defamiliarization** (Q5; novel expression), and **Language Artistry** (Q6; economy/rhythm/pacing). Q2–Q4 are grounded in reader-response formalisms that model perceived intent and explanatory justification as separable reception dimensions (Mire et al., 2025); Q5 draws on defamiliarization as a literary technique (Shklovsky, 1965); and Q6 aligns with creative-writing assessment rubrics and stylistic accounts of comic timing in prose (Vaezi and Rezaei, 2019; Haines, 2024).

Social Framing & Downstream Impact. To capture social positioning, we adapt the **Humor Styles Questionnaire** (Q7–Q10: Affiliative, Self-enhancing, Aggressive, Self-defeating) (Martin et al., 2003). Finally, we evaluate downstream reception via: **Value Judgment Pressure** (Q11) (Mire et al., 2025), **Memorability** (Q12) (Gopi and Madan, 2024), **Share Willingness** (Q13) (Norman and Russell, 2006), and **Social/Task Attraction** (Q14–Q15) (McCroskey and McCain, 1974).

4.2 Human Evaluation Protocol

Raters. We recruited dedicated raters who completed the full annotation workload. We used dedicated raters (instead of open crowdworkers) as our

diagnostic metrics target writing craft and mechanisms (e.g., intent clarity and explainable turns) that benefit from a shared rubric interpretation.

Task and blinding. For each matched pair, raters read the topic prompt followed by two anonymized texts (A/B), which were randomized independently per item. Meanwhile, all items (item = topic × performer × round) were **shuffled**, and raters saw items from **non-consecutive rounds** in a fully mixed order. This also reduces learning and fatigue artifacts that could otherwise correlate with rounds.

4.3 Inter-rater reliability.

Five raters evaluated each paired comparison, providing (i) a binary preference (Q0: prefer A vs. B) and (ii) Likert ratings (Q1–Q15, 1–5) for both text A and text B. Agreement on Q0 was fair (Fleiss' κ=0.237, 95% CI [0.171, 0.299]; Gwet's AC1 =0.253, [0.188, 0.321]; N=249) (Fleiss, 1971; Gwet, 2008). For Likert items, we analyzed the consistency of per-rater difference scores using the average-measures ICC(3,5); reliability was substantially higher (ICC(3,5) =0.710, [0.640, 0.765]; N=241) (Shrout and Fleiss, 1979; Koo and Li, 2016). Full details and subscale reliabilities are provided in Appendix D.

Preference votes compress multiple criteria (e.g., humor taste, perceived offensiveness, and personal norms) into a single forced choice, making them inherently subjective and prone to split decisions when paired texts are close. In our data, only 29.2% (73/250) of instances are unanimous (5–0) in favor of DISCUSSION, and nearly half are decided by narrow margins: 26.4% (66/250) are 3–2 "wins" and 20.8% (52/250) are 2–3 "losses;" the remaining cases are 4–1 wins (20.0%, 50/250) and 1–4 losses (3.6%, 9/250). Such frequent near-ties naturally depress chance-corrected agreement on Q0, so we treat Q0 as a *supporting* signal instead of a primary outcome. In contrast, the Likert items provide more diagnostic and specific judgments. Accordingly, our main analyses rely on Likert-based difference.

5 Results

Table 1 summarizes per-item human evaluation (Q0–Q15). Across paired instances, Discussion-enabled outputs are preferred more often than Baseline (Q0) and show consistent improvements on Craft/Clarity (Q1–Q6) and Social Response (Q12–Q15). However, humor-style items (Q7–Q10) are not monotonic "higher-is-better" outcomes: increases can reflect either benign/affiliative strengthening or harmful/maladaptive intensification.

Paired estimation and confidence intervals. Because A/B presentation is randomized per instance, we first map each rater's A/B ratings back to condition identity using `A_System/B_System`. For each Likert item $q \in \{Q1, \ldots, Q15\}$ and paired instance i (topic×performer×round; N=250), each rater r yields a paired difference $\delta_{i,r,q} = y_{i,r,q}(\text{Discussion}) - y_{i,r,q}(\text{Baseline})$. To respect repeated measures, we average within instance across raters, $\Delta_{i,q} = \frac{1}{|R_i|}\sum_{r\in R_i}\delta_{i,r,q}$, and report mean effects across instances. All 95% CIs for Q1–Q15 in Table 1 are clustered-bootstrap percentile intervals obtained by resampling instances (B=20,000); recorded zeros are treated as missing. For Q0, we report individual vote shares and the instance-level majority-win rate with a Wilson 95% CI.

5.1 Overall Gains on Primary Outcomes

Discussion wins the instance-level majority vote in 75.6% of cases (189/250; Wilson 95% CI [69.9, 80.5]) and receives 70.1% of individual votes (876/1249). Aggregating item-level differences into the two primary profiles, Discussion yields clear gains on: Craft/Clarity (Q1–Q6) $\overline{\Delta}$=0.440 and Social Response (Q12–Q15) $\overline{\Delta}$=0.422. At the item level (Table 1), all Craft/Clarity and Social Response items shift positively, with large improvements on Q1 (Immediate Amusement), Q4 (Justified Landing), Q12 (Memorability), and Q15 (Task Attraction).

Humor style direction via HarmShift. We decompose humor styles into benign/affiliative components (Q7,Q8) and harmful/maladaptive components (Q9,Q10). For each instance i, we define:

$$\Delta\text{Craft}_i = \tfrac{1}{5}\sum_{q=2}^{6}\Delta Q_{i,q},$$

$$\Delta\text{Downstream}_i = \tfrac{1}{4}\sum_{q=12}^{15}\Delta Q_{i,q},$$

$$\text{HarmShift}_i = \tfrac{1}{2}(\Delta Q_{i,9} + \Delta Q_{i,10}) - \tfrac{1}{2}(\Delta Q_{i,7} + \Delta Q_{i,8}),$$

$$\Delta\text{Pref}_i = \text{PrefShare}_i - 0.5.$$

where $\text{PrefShare}_i \in [0, 1]$ is the fraction of rater votes preferring Discussion. This avoids the ambiguity that arises when all four style items increase simultaneously: HarmShift > 0 indicates a net shift toward harmful/maladaptive style, even if benign styles also strengthen.

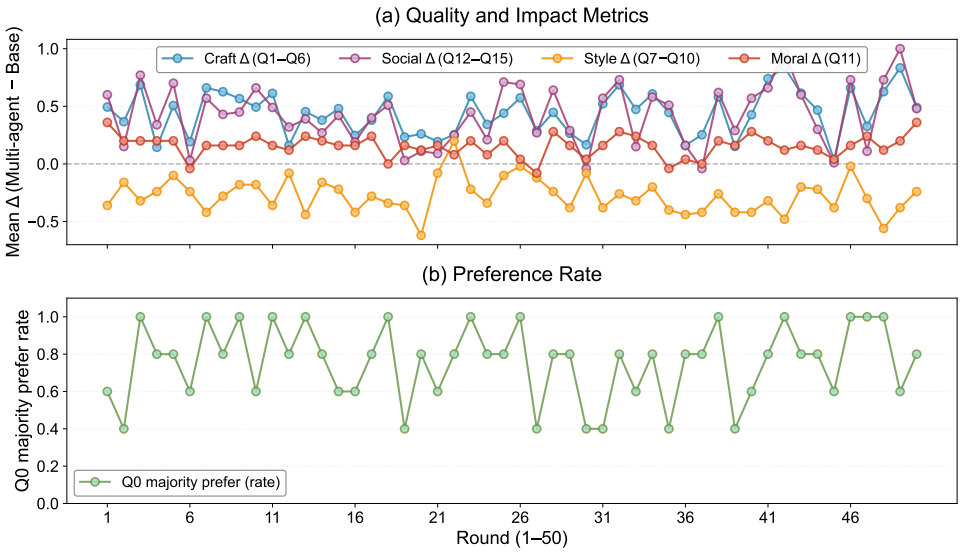

Figure 4: **Round-to-round dynamics.** (a) Round-level mean differences Δ = Discussion − Baseline for Craft/Clarity (Q1–Q6), Social Response (Q12–Q15), and Moral Pressure (Q11). We report Humor Style direction with *HarmShift* = $\text{mean}(\Delta Q9, \Delta Q10) - \text{mean}(\Delta Q7, \Delta Q8)$ (higher = more harmful shift). (b) The instance-level Q0 majority preference rate for Discussion in each round.

5.2 Stability across Rounds and Performers

Figure 4 shows round-level mean differences and the Q0 majority-win rate. Craft/Clarity and Social Response advantages remain mostly positive across rounds, while preference varies by topic, consistent with prompt-dependent difficulty and varying proximity between paired outputs. Aggregating by performer persona yields the same qualitative pattern (Appendix E); between-performer differences are not statistically reliable for Craft, Social, or HarmShift (all $p > 0.1$), suggesting the overall gains are not driven by a single performer.

5.3 Benefit-Safety Tradeoff

We construct a composite *Benefit* and *Safety* score per instance (Figure 5). Each point is one paired instance (topic×performer×round). Let $z(\cdot)$ denote z-scoring across instances. We define:

$$\text{Benefit}_i = \tfrac{1}{4}\Big(z(\Delta Q_{i,1}) + z(\Delta\text{Craft}_i) + z(\Delta\text{Downstream}_i) + z(\Delta\text{Pref}_i)\Big),$$

$$\text{Safety}_i = -\tfrac{1}{2}\Big(z(\Delta Q_{i,11}) + z(\text{HarmShift}_i)\Big).$$

Here $\Delta\text{PrefShare}_i$ is the rater preference share for Discussion centered at 0.5 (ties at 0). The crosshairs at $(0, 0)$ mark the dataset means (z=0), so the upper-right quadrant represents instances above average on both overall gains and safety. In our data, 57/250 instances (22.8%) fall in this "win–win" quadrant. We also highlight Pareto-efficient points (6/250, 2.4%), which are not dominated by any other instance in the joint objective of maximizing Benefit and Safety. Benefit and Safety are only weakly correlated overall (Spearman $\rho = -0.046$, $p = 0.472$), indicating heterogeneous tradeoffs rather than a single monotonic coupling.

ID	Metric	Scale	Discuss.	Base	Δ	95% CI
Q0	Preference	A/B	70.1%	29.9%	75.6%	[69.9, 80.5]
Outcome & Mechanism/Craft Profile						
Q1	Immediate Amusement	1–5	2.85	2.33	0.52	[0.44, 0.59]
Q2	Reframing / Insight	1–5	2.92	2.47	0.45	[0.38, 0.51]
Q3	Intent Clarity	1–5	3.34	3.06	0.27	[0.21, 0.33]
Q4	Justified Landing	1–5	3.12	2.63	0.49	[0.42, 0.56]
Q5	Defamiliarization	1–5	2.86	2.39	0.46	[0.40, 0.53]
Q6	Language Artistry	1–5	3.04	2.58	0.45	[0.38, 0.53]
Humor Style (HSQ-adapted)						
Q7	Affiliative	1–5	2.59	2.51	0.08	[0.03, 0.13]
Q8	Self-enhancing	1–5	1.90	1.85	0.05	[0.01, 0.09]
Q9	Aggressive	1–5	2.69	2.26	0.42	[0.36, 0.49]
Q10	Self-defeating	1–5	2.18	1.93	0.25	[0.20, 0.30]
Social Framing & Downstream Impact						
Q11	Value Judgment Pressure	1–5	1.76	1.61	0.16	[0.12, 0.19]
Q12	Memorability	1–5	2.81	2.34	0.46	[0.38, 0.54]
Q13	Share Willingness	1–5	2.48	2.05	0.44	[0.36, 0.51]
Q14	Social Attraction	1–5	2.65	2.35	0.30	[0.23, 0.37]
Q15	Task Attraction	1–5	2.79	2.30	0.49	[0.42, 0.56]

Table 1: Per-item human evaluation results. For Q1–Q15, Δ = Discussion − Baseline (instance-level; N=250). For Q0, "Discuss."/"Base" are individual vote shares and Δ is the majority-win rate (189/250) with Wilson 95% CI.

5.4 Interpretation

To interpret why multi-agent outperforms baseline across *all* dimensions, we qualitatively analyze the writing changes it induces. Across topics, multi-agent discussion tends to produce a tightly bundled set of rhetorical moves (early premise commitment,

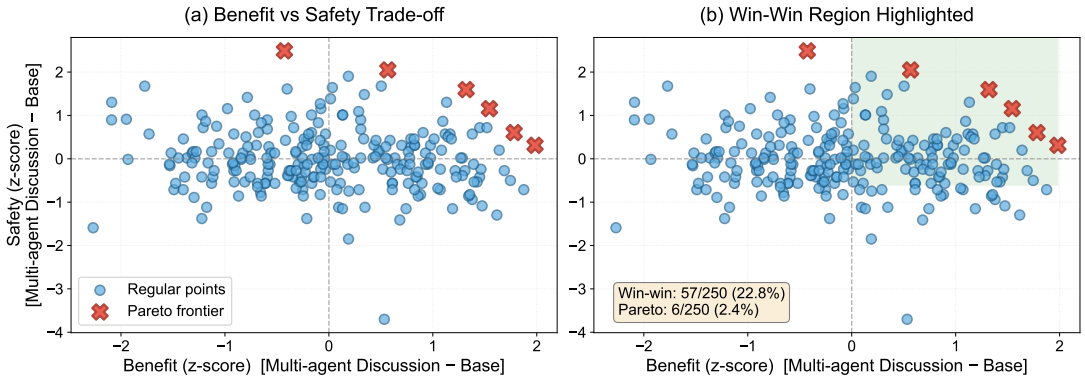

Figure 5: **Benefit–safety tradeoff.** Each point is a paired instance (topic×performer×round). Benefit (x-axis; z-scored, higher is better) averages gains in amusement (Q1), craft (Q2–Q6), downstream impact (Q12–Q15), and centered preference share ($\mathrm{PrefShare} - 0.5$). Safety (y-axis; z-scored, higher is better) is the negative mean of moral/value-judgment pressure shift (Q11) and style-direction shift *HarmShift*. Dashed crosshairs mark dataset means (z=0). Red X marks indicate Pareto-efficient instances; panel (b) highlights the win–win quadrant (Benefit≥ 0, Safety≥ 0).

sustained personification, one-axis escalation, and decisive endings) that raises multiple ratings simultaneously including both benign and risky style dimensions. Appendix F grounds this account with verbatim multi-agent vs. baseline excerpts and brief mechanistic annotations.

6 Discussion and Conclusion

The key findings are summarized below. First, in *stand-up comedy monologue generation*, adding a *broadcast community discussion* leads to better humor outputs than a no-discussion baseline. The discussion-enabled system achieves gains in Craft/Clarity (Δ=0.440) and Social Response (Δ=0.422), with 75.6% majority preference. These effects are consistent across rounds and performer personas, suggesting that a *persistent reception stream* can be operationalized as *retrievable social memory* to condition later contexts in multi-round creative writing.

Second, craft improvements come with tradeoffs. The coupling between craft gains and aggressive humor ($\rho = 0.289$) suggests community discussion may encourage edgier comedic strategies, as such material generates more distinctive reception signals. Only 20–25% of instances achieve craft gains without increases in these risk-associated dimensions, raising questions for deployment contexts where content moderation matters.

Third, our diagnostic evaluation protocol, pairing a forced-choice preference vote (Q0) with multi-dimensional rubric ratings (Q1–Q15), supports a more informative assessment than a single binary judgment. As expected for humor, overall preference exhibits only modest rater agreement ($\kappa = 0.237$), reflecting that a forced choice compresses multiple criteria and personal taste into one decision. Crucially, the rubric-based *difference scores* are substantially more consistent across raters (ICC(3,5)$= 0.710$), indicating that annotators reliably agree on *which specific qualities improved* even when they may disagree on an overall winner. These results validate our evaluation metrics that relative comparisons can be assessed reliably despite humor's inherent subjectivity.

The multi-agent system we introduce generalizes beyond stand-up comedy to other audience-oriented creative domains. Fiction writing communities, collaborative screenwriting, and persuasive content creation all feature public reception streams that shape iterative production. Although the core architecture remains unchanged, adaptation to specific domains occurs primarily at the reception abstraction layer. This is because different fields emphasize distinct units of highly informative feedback, such as narrative coherence and character voice in fiction, pacing and scene transitions in screenwriting, or perceived credibility and alignment of stance in persuasive writing. Consequently, the memory filter and retrieval criteria can be modified to prioritize reception signals pertinent to the domain while maintaining the established bounded social memory interface and retrieval mechanism based on embeddings. The core insight that broadcast community discussion provides implicit supervision for creative improvement suggests broader applications in educational writing environments, collaborative design platforms, and social media content optimization.

Limitations

This paper presents several opportunities for future research. First, all agents in our simulation are driven by GPT-4o-mini. While this ensures internal consistency, it limits the analysis of other LLMs. Future work could examine whether the observed effects of community discussion replicate across diverse model families (e.g., Claude, Llama, Gemini) and model scales, as humor generation capabilities may differ substantially between different models and training configurations.

Second, our evaluation is based on 50 rounds yielding 250 monologues per condition with a fixed topic list. Longer simulation horizons and more diverse topic distributions can reveal additional dynamics in how community feedback shapes comedic output over extended periods. The topics selected may also inadvertently favor certain comedic styles or performer personas over others, potentially introducing biases into the comparative results. Future studies could expand the topic pool and incorporate user-generated or culturally varied prompts to improve generalizability.

Lastly, human evaluation of humor is inherently subjective and culturally situated. Our annotator pool may not fully reflect universal comedic preferences across different cultures and age groups, and evaluating decontextualized monologues outside their simulated community setting may not fully capture the social dynamics we aim to study. Future research could involve larger, more diverse designs with broader demographic representations, which would strengthen the robustness of our findings.

Ethical Considerations

Stand-up comedy frequently engages with sensitive topics including social taboos and controversial viewpoints. While our performer personas are designed with diverse comedic styles, the simulation may generate content that some audiences find offensive or inappropriate. We implement persona-based guidelines but do not employ additional content filtering to preserve ecological validity. Additionally, the simulated community feedback mechanisms can, if deployed in real systems, amplify certain comedic styles while marginalizing others, potentially creating homogenization effects. Researchers deploying similar systems should carefully consider content moderation strategies and potential biases in feedback loops appropriate to their specific use cases and target audiences.

We will release the sandbox configuration and code under the MIT License, and release the generated artifacts (paired monologues, reconstructed discussion threads, and event logs) under CC BY-NC 4.0 for research and reproducibility purposes. We emphasize that these artifacts are intended for research use rather than deployment as an end-user comedy system, and that any reproduction of our pipeline should comply with the access conditions of the underlying LLM/API.

Prior to annotation, raters were informed of the possibility of exposure to offensive or sensitive content typical of stand-up comedy; participation was voluntary, and raters could skip any questions or stop at any time. We collected only rubric ratings and preference judgments, and we report results in aggregate and do not collect or release personally identifying information about raters. To support interpretation of the human evaluation, we characterize the annotator pool at a high level: raters were with sufficient English proficiency to assess long-form comedic text and familiarity with the evaluation rubric.

Acknowledgments

<Omitted for double-blind review>

References

Miriam Amin and Manuel Burghardt. 2020. A survey on approaches to computational humor generation. In *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 29–41, Online. International Committee on Computational Linguistics.

Yi Cao, Jiahao Cao, Yubo Hou, and Li-Jun Ji. 2025. How humorous is ai? exploring chatgpt's role in humor generation and human-ai interaction. *Computers in Human Behavior Reports*, 20:100807.

Eric Chamoun, Michael Schlichtkrull, and Andreas Vlachos. 2024. Automated focused feedback generation for scientific writing assistance. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9742–9763, Bangkok, Thailand. Association for Computational Linguistics.

Ruijia Cheng and Jenna Frens. 2022. Feedback exchange and online affinity: A case study of online fanfiction writers. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Alessio Cocchieri, Luca Ragazzi, Paolo Italiani, Giuseppe Tagliavini, and Gianluca Moro. 2025. "what do you call a dog that is incontrovertibly true?

dogma": Testing LLM generalization through humor. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22922–22937, Vienna, Austria. Association for Computational Linguistics.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Chen Gao, Xiaochong Lan, Nian Li, Yuan Yuan, Jingtao Ding, Zhilun Zhou, Fengli Xu, and Yong Li. 2024. Large language models empowered agent-based modeling and simulation: a survey and perspectives. *Humanities and Social Sciences Communications*, 11:1259.

Yashoda Gopi and Christopher R. Madan. 2024. Subjective memory measures: Metamemory questionnaires currently in use. *Quarterly Journal of Experimental Psychology*, 77(5):924–942.

Drew Gorenz and Norbert Schwarz. 2024. How funny is chatgpt? a comparison of human- and a.i.-produced jokes. *PLOS ONE*, 19(7):e0305364.

Qingyu Guo, Chao Zhang, Hanfang Lyu, Zhenhui Peng, and Xiaojuan Ma. 2023. What makes creators engage with online critiques? understanding the role of artifacts' creation stage, characteristics of community comments, and their interactions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Kilem L. Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Alice Haines. 2024. Comic timing in prose fiction. *Journal of Literary Semantics*, 53(2):93–109.

Senyu Han, Lu Chen, Li-Min Lin, Zhengshan Xu, and Kai Yu. 2024. IBSEN: Director-actor agent collaboration for controllable and interactive drama script generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1607–1619, Bangkok, Thailand. Association for Computational Linguistics.

Fantine Huot, Reinald Kim Amplayo, Jennimaria Palomaki, Alice Shoshana Jakobovits, Elizabeth Clark, and Mirella Lapata. 2025. Agents' room: Narrative generation through multi-step collaboration. In *The Thirteenth International Conference on Learning Representations (ICLR 2025), Singapore, April 24–28, 2025*. OpenReview.net.

Terry K. Koo and Mae Y. Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2):155–163.

Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, Paul Liang Liang, Luis Alonso, and Kent Larson. 2025. Simulating society requires simulating thought. *Preprint*, arXiv:2506.06958. NeurIPS 2025 (Position Paper Track).

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, (140):1–55.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Tyler Loakman, William Thorne, and Chenghua Lin. 2025. Who's laughing now? an overview of computational humour generation and explanation. In *Proceedings of the 18th International Natural Language Generation Conference*, pages 780–794, Hanoi, Vietnam. Association for Computational Linguistics.

Rod A. Martin, Patricia Puhlik-Doris, Gwen Larsen, Jeanette Gray, and Kelly Weir. 2003. Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality*, 37(1):48–75.

James C. McCroskey and Thomas A. McCain. 1974. The measurement of interpersonal attraction. *Speech Monographs*, 41(3):261–266.

Qirui Mi, Mengyue Yang, Xiangning Yu, Zhiyu Zhao, Cheng Deng, Bo An, Haifeng Zhang, Xu Chen, and Jun Wang. 2025. MF-LLM: Simulating population decision dynamics via a mean-field large language model framework. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Joel Mire, Maria Antoniak, Steven R. Wilson, Zexin Ma, Achyutarama R. Ganti, Andrew Piper, and Maarten Sap. 2025. Social story frames: Contextual reasoning about narrative intent and reception. *Preprint*, arXiv:2512.15925.

Piotr Mirowski, Kory Mathewson, and Boyd Branch. 2025. The theater stage as laboratory: Review of real-time comedy LLM systems for live performance. In *Proceedings of the 1st Workshop on Computational Humor (CHum)*, pages 88–95, Online. Association for Computational Linguistics.

Sheshera Mysore, Zhuoran Lu, Mengting Wan, Longqi Yang, Bahareh Sarrafzadeh, Steve Menezes, Tina Baghaee, Emmanuel Barajas Gonzalez, Jennifer Neville, and Tara Safavi. 2024. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers. In *Proceedings of the 1st Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pages 198–219, Miami, Florida, USA. Association for Computational Linguistics.

Reuben Narad, Siddharth Suresh, Jiayi Chen, Pine S. L. Dysart-Bricken, Bob Mankoff, Robert Nowak, Jifan Zhang, and Lalit Jain. 2025. Which llms get the joke? probing non-stem reasoning abilities with humorbench. *arXiv preprint arXiv:2507.21476*.

Andrew T. Norman and Cristel A. Russell. 2006. The pass-along effect: Investigating word-of-mouth effects on online survey procedures. *Journal of Computer-Mediated Communication*, 11(4):1085–1103.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23, New York, NY, USA. Association for Computing Machinery.

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. *Preprint*, arXiv:2502.08691.

Maximilian Puelma Touzel, Sneheel Sarangi, Gayatri Krishnakumar, Busra Tugce Gurbuz, Austin Welch, Zachary Yang, Andreea Musulan, Hao Yu, Ethan Kosak-Hine, Tom Gibbs, Camille Thibault, Reihaneh Rabbany, Jean-François Godbout, Dan Zhao, and Kellin Pelrine. 2025. Sandboxsocial: A sandbox for social media using multimodal ai agents. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 11100–11103. International Joint Conferences on Artificial Intelligence Organization. Demo Track.

Kristina Radivojevic, Matthew Chou, Karla Badillo-Urquiola, and Paul Brenner. 2024. Human perception of LLM-generated text content in social media environments. *Preprint*, arXiv:2409.06653.

Yiting Ran, Xintao Wang, Tian Qiu, Jiaqing Liang, Yanghua Xiao, and Deqing Yang. 2025. BOOKWORLD: From novels to interactive agent societies for story creation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15898–15912, Vienna, Austria. Association for Computational Linguistics.

Victor Raskin. 1979. Semantic mechanisms of humor. In *Proceedings of the Fifth Annual Meeting of the Berkeley Linguistics Society (BLS 5)*, pages 325–335.

Sahithya Ravi, Patrick Huber, Akshat Shrivastava, Vered Shwartz, and Arash Einolghozati. 2024. Small but funny: A feedback-driven approach to humor distillation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13078–13090, Bangkok, Thailand. Association for Computational Linguistics.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: language agents with verbal reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Viktor Shklovsky. 1965. Art as technique. In Lee T. Lemon and Marion J. Reis, editors, *Russian Formalist Criticism: Four Essays*, pages 3–24. University of Nebraska Press, Lincoln, NE. Originally published 1917.

Patrick E. Shrout and Joseph L. Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Maryam Vaezi and Shahla Rezaei. 2019. Development of a rubric for evaluating creative writing: A multiphase research. *New Writing*, 16(3):303–317.

Lei Wang, Jingsen Zhang, Hao Yang, Zhi-Yuan Chen, Jiakai Tang, Zeyu Zhang, Xu Chen, Yankai Lin, Hao Sun, Ruihua Song, Xin Zhao, Jun Xu, Zhicheng Dou, Jun Wang, and Ji-Rong Wen. 2025. User behavior simulation with large language model-based agents. *ACM Trans. Inf. Syst.*, 43(2).

Noah Wang, Z.y. Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, Man Zhang, Zhaoxiang Zhang, Wanli Ouyang, Ke Xu, Wenhao Huang, Jie Fu, and Junran Peng. 2024a. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777, Bangkok, Thailand. Association for Computational Linguistics.

Qineng Wang, Zihao Wang, Ying Su, Hanghang Tong, and Yangqiu Song. 2024b. Rethinking the bounds of LLM reasoning: Are multi-agent discussions the key? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6106–6131, Bangkok, Thailand. Association for Computational Linguistics.

Tian Yu, Ken Shi, Zixin Zhao, and Gerald Penn. 2025. Multi-agent based character simulation for story writing. In *Proceedings of the Fourth Workshop on Intelligent and Interactive Writing Assistants (In2Writing 2025)*, pages 87–108, Albuquerque, New Mexico, US. Association for Computational Linguistics.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, and Jiaxuan You. 2025. MultiAgentBench : Evaluating the collaboration and competition of LLM agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8580–8622, Vienna, Austria. Association for Computational Linguistics.

A Event Log and Thread Assignment Rule

Event log. We store the full simulation trace as an event log $\mathcal{E} = \{e_1, \ldots, e_N\}$.
Each event e is a JSON object with core fields:

`type:`	$\in$ {`moderator_topic`, `performance`, `critic_review`, `free_dialogue`}.
`round:`	integer round index t.
`timestamp:`	ISO 8601 string.
`author:`	agent name.
`content:`	textual payload (or structured list for performances).
`mentions:`	list of referenced agent names.
`replyTo:`	optional primary target agent name for replies.
`replyToThreadId:`	optional explicit thread id selected by the agent.
`thread_id:`	UUID identifying a discussion thread.
`parent_id:`	optional; for replies equals the root `thread_id`; None for thread starters.

Thread assignment rule. We assign `thread_id` using the following precedence:

1) If `replyToThreadId` present:	`thread_id` ← `replyToThreadId`; `parent_id` ← `thread_id`.
2) Else if `replyTo` present:	Find the most recent prior event authored by `replyTo` within the same round, inherit its `thread_id`, and set `parent_id` accordingly.
3) Otherwise:	Start a new thread with a fresh UUID and set `parent_id=None`.

B Human Evaluation Metric Items

Participants read two texts (A and B). For the paired preference question (Q0), participants select which text they prefer overall. For the per-text ratings (Q1–Q15), participants provide a 1–5 Likert-type scale rating (Likert, 1932) for each text separately, where 1 = Strongly disagree / Not at all, and 5 = Strongly agree / Very much (Table 2).

C Full Persona Text for All Agents (N=35)

Below is the persona text for all Performer Agents.

```
Emma - Performer: Emma is a stand-up comedian
    known for her dark humor and social satire.
    Born into an urban middle-class family, she
    received a solid education, majoring in
    scriptwriting with a minor in phenomenology.
     During university, she befriended Luna, a
    renowned stand-up critic whose influence
    helped shape her artistic sensibilities.
    Deeply engaged with feminist issues and the
    marginal narratives of minority groups, Emma
    's performances often challenge conventions
    and social taboos. Her comedic style leans
    toward experimental and artistic expression,
     constantly pushing the boundaries of what
    stand-up comedy can be. Outside the stage,
    she enjoys opera, musical theatre, and
    visiting art exhibitions. In every
    performance, Emma experiments with new joke
    structures and delivery methods, carefully
    studying audience reactions and critical
    feedback to refine her creative strategy.
    Her goal is to make the audience both laugh
    and think - to balance humor with
    intellectual and aesthetic depth while
    continually redefining her personal comedic
    voice.

Max - Performer: Max is a stand-up comedian with
     a background in advertising and a deep
    sensitivity to audience reception and
    commercial value. Living single in a big
    city, he crafts comedy that resonates with
    mainstream audiences - light, fast-paced,
    and filled with familiar cultural references
    . His humor often draws on everyday themes
    such as romantic relationships, parenting,
    workplace absurdities, and the influence of
    new technologies. Max thrives on audience
    interaction and high joke density, ensuring
    his shows remain accessible and energetic.
    An avid follower of celebrity variety shows
    and sports, he has a sharp, outgoing
    personality and a keen sense of timing. Max
    skillfully capitalizes on trending topics,
    especially gender debates and social
    tensions, using them to spark laughter and
    conversation alike. A master of reading both
     live audiences and social media analytics,
    he adapts his material in real time to
    maximize engagement. His ultimate goal is to
     achieve viral impact - to turn laughter
    into attention while maintaining
    professional polish and creative control.

Alice - Performer: Alice is a stand-up comedian
    who previously worked in a major tech
    company specializing in artificial
    intelligence. Her background gives her deep
    insight into technological development, AI
    ethics, and the social consequences of
    automation. In her performances, Alice
    transforms weighty subjects such as
    algorithmic bias, data privacy, and the
    erosion of jobs into sharp, ironic humor.
    Her style carries an undertone of
    disillusioned clarity - the sense of 'seeing
     through everything but being powerless to
    change it' - balancing cynicism with wit.
    Through her comedy, Alice aims to disrupt
    both blind faith in and irrational fear of
    technology. She invites audiences to remain
    alert about their data, labor, and autonomy,
     wielding humor as a subtle but incisive
    instrument against the illusions of
    technological utopianism.

Leo - Performer: Leo is a stand-up comedian
    shaped by his grassroots upbringing in a
    union family, where discussions of labor
    movements and collective struggle were part
    of daily life. A self-taught reader of
    social theory and classic leftist texts, he
    distills complex ideas into humor that is
```

Table 2: Human evaluation metrics. Q0 is a paired preference item; Q1–Q15 are rated on a 1–5 Likert-type scale.

ID	Metric	Description
Q0	Preference (A/B)	Overall, which text do you prefer? (Choose one: A or B.)
Outcome & Mechanism/Craft Profile		
Q1	Immediate Amusement	Did this text make you laugh?
Q2	Reframing / Insight	This text gives me a reframing/insight or makes me more sensitive to an experience.
Q3	Perceived Intent Clarity	I can tell what this text is trying to accomplish (e.g., amuse, vent, self-expression, persuade, empathize).
Q4	Justified Landing	After reading this text, I can look back and point to cues that support how the turn lands. The turn feels justified and coherent.
Q5	Defamiliarization	This text uses language/imagery/rhetoric in a fresh way that makes me see something familiar differently.
Q6	Language Artistry	This text's sentence economy, rhythm, and keyword choices effectively serve the punch/impact. There is little unnecessary filler.
Humor Style (adapted from Humor Styles Questionnaire)		
Q7	Affiliative	The use of humor to enhance relationships with others.
Q8	Self-enhancing	The use of humor to enhance the self.
Q9	Aggressive	The use of humor to enhance the self at the expense of others.
Q10	Self-defeating	The use of humor to enhance relationships at the expense of the self.
Social Framing & Downstream Impact		
Q11	Value Judgment Pressure	While reading this text, I felt pressure to make a strong value/moral judgment (e.g., "Is this acceptable?" "Which side am I on?").
Q12	Memorability	After finishing this text, how much of it can you remember without re-reading (e.g., key lines, images, or the main turn)?
Q13	Share Willingness	How willing would you be to share this text with a friend (e.g., forward it, repost it, or send it in a group chat)?
Q14	Social Attraction	After reading this text, the "speaker" feels likable/cute, and I would be willing to keep listening or be friends.
Q15	Task Attraction	After reading this text, the "speaker" feels skilled, and I would trust them to handle creative writing.

Note: *Scale anchors: 1 = Strongly disagree / Not at all; 2 = Disagree / Slightly; 3 = Neutral / Somewhat; 4 = Agree / Quite a bit; 5 = Strongly agree / Very much.*

```
    sharp, grounded, and accessible. Leo
    frequently invokes themes of wealth
    inequality, class immobility, worker
    exploitation, and systemic injustice, using
    class conflict as a comedic lens on issues
    like housing prices and the gig economy. His
     routines often resemble 'class analysis
    lectures,' delivered with deadpan
    seriousness before twisting into absurdist
    punchlines that land with both humor and
    impact. A devoted anime fan, Leo tailors his
     material according to audience reactions,
    tapping into shared frustrations and social
    pain points. His aim is to provoke laughter
    that burns with recognition - to make
    audiences laugh in anger and reflect
    afterward, tracing personal discontent back
    to structural causes.

Richard - Performer: Richard is a stand-up
    comedian with a background in history,
    having specialized in Roman philosophy
    before taking an office job he eventually
    came to despise. His comedy dissects the
    absurdities of corporate culture - from
    management's PUA-style manipulation and
    euphemistic layoffs to the hollow jargon of
    'team-building' and 'innovation.' His
    signature style is razor-sharp satire: he
    might deliver a line like, '996 isn't
    exploitation - it's a Self-Driven Talent
    Potential Unleash Program,' in a monotone
    PowerPoint voice, only to punctuate it with
    a brutally honest punchline. By mimicking
    executives on dull Zoom calls, Richard
    exposes the hypocrisy and emptiness of
    corporate rhetoric. To him, corporate jargon
     is the new 'spiritual opium,' and he
    positions himself as an insider
    whistleblower, revealing every layer of
    linguistic and psychological control that
    sustains workplace misery. Constantly
    updating his 'corporate bullshit dictionary
    ,' Richard keeps his material in sync with
    the latest buzzwords and trends. His mission
     is clear: to act as a reverse consultant -
    dismantling corporate power structures with
    humor as precise as it is merciless.
```

Below is the persona text for all Critic Agents.

```
Luna - Critic: Luna is a well-known stand-up
    comedy critic and freelance writer. She
    majored in phenomenology and often cites
    philosophy, sociology, and theater theory in
     her analyses. Luna frequently interprets
    stand-up performances within larger social
    contexts. She has a rational reviewing style
    , a high sensitivity to social issues, and
    takes a clear stance in her critiques. She
    personally dislikes awkward or lowbrow humor
    . Her critique style is sharp and
    provocative; she delivers harsh criticism
```

```
    when a performance does not meet her
    standards. She supports performers who push
    boundaries and experiment, even if it risks
    alienating mainstream audiences. Her goal is
     to expand the boundaries of stand-up comedy
    , encourage innovative and experimental
    expression, and direct audiences' attention
    to non-mainstream culture and marginalized
    social issues. In her free time, she enjoys
    attending operas and musicals with Emma, as
    well as visiting art exhibitions.

Ethan - Critic: Ethan is a professional stand-up
     comedy critic and freelance writer. He is
    deeply interested in analyzing the mechanics
     of humor, deconstructing jokes academically
     in terms of structure, emotional mechanisms
    , and cultural significance. Sometimes he
    gets 'Lacan-obsessed,' attempting to
    interpret everything through Lacanian theory
    . He is attentive to social issues and
    politics, often exploring why certain jokes
    are embraced by specific social groups,
    including the political or class context
    behind them. He enjoys playing all kinds of
    games, including anime-style games and MOBAs
    , which is how he met Leo.

Clara - Critic: Clara is a theater critic with a
     literary background, specializing in
    dramatic structure, and also works part-time
     as a stand-up comedy reviewer. She focuses
    on narrative lines, emotional transitions,
    and language details in performances. She is
     skilled at analyzing joke structure and
    techniques, such as setup, punchline, and
    callback pacing. Her reviews are generally
    gentler than Luna or Ethan's, though she
    still offers criticism. She sometimes
    examines performances from the perspective
    of commercial value and dissemination
    strategy. Overall, she supports generating
    topical conflicts to stimulate audience
    engagement.
```

The persona text for all *Audience Agents*:

```
Sophia - Audienc: Third-year psychology student.
     Outgoing, uses humor to relieve study
    stress. Enjoys mainstream pop culture,
    romantic comedies, and relatable college
    life jokes. Very engaged with social media
    trends. Friend is Iris.

Iris - Audience: Third-year sociology student
    and barista. Appreciates short,
    philosophical jokes that comment on cultural
     meanings and societal structure. Discusses
    the deeper meaning of humor with her friend
    Sophia.

Daniel - Audience: Product manager at a major
    tech company. Rational and highly picky
    about delivery and structure. Observes joke
    timing, pacing, and logical flow critically.
     Often finds flaws in poorly constructed
    material.

Lily - Audience: Works in parcel logistics.
    Outgoing and direct. Loves life-based humor,
     physical comedy, and is a big fan of anime,
     often referencing it. Her humor taste is
    often the opposite of her husband Daniel's.

Jimmy - Audience: Introverted programmer. His
    world revolves around code and tech. He
    enjoys humor that references coding culture,
     AI, and geek culture. Is highly sensitive
    to crude or overly offensive material.

Olivia - Audience: Graphic designer by trade.
    She values artistic innovation, visual
    pacing, and creative delivery in comedy. She
     treats stand-up like a piece of visual art,
     observing the 'composition' of the joke.

Grace - Audience: A new office worker, still
    energetic and curious about the corporate
    world. She prefers fast-paced, interactive
    humor, especially jokes that focus on the
    absurdity of the workplace and young adult
    struggles.

Jason - Audience: Factory worker. Frank, direct,
     and strongly affected by social injustice
    and class disparities. He appreciates comedy
     that speaks truth to power and exposes
    wealth inequality. A close friend and
    colleague of Tom.

Tom - Audience: New factory worker, highly
    sensitive to labor rights and class issues
    due to his new job environment. He prefers
    comedy that is socially critical and
    satirical. Discusses social analysis with
    his friend Jason.

Mark - Audience : Professional software engineer
     and husband of performer Alice. His taste
    is centered on logical humor, life
    observations, and emotional honesty, rather
    than pure tech humor. Supports his wife but
    judges objectively.

Paul - Audience: Administrative assistant.
    Steady and calm. Enjoys workplace satire and
     life humor. He was a former colleague and
    remains a friend of performer Richard, often
     understanding Richard's corporate jokes
    deeply.

Julia - Audience: High school student, sensitive
     and literary-minded. Prefers jokes with
    literary, cultural, or subtle psychological
    humor. Finds deep meaning in wordplay.
    Sister of critic Clara, often influenced by
    her literary background.

Mia - Audience: Elementary school teacher.
    Prefers warmth, gentle life observation, and
     emotionally detailed humor. She seeks
    connection and humanity in comedy. Sister of
     performer Max, she is generally supportive
    but values sincerity.

Victor - Audience: From a business family, very
    practical and direct. He prefers satire
    emphasizing personal effort, career success,
     and practical life humor. He actively
    dislikes over-politicized social conflict
```

```
    humor, viewing it as unproductive.

Ryan - Audience: Fitness coach, energetic and
    lively. He enjoys male-perspective humor and
     jokes that play on gender differences or
    conflicts, often finding humor in
    traditional gender stereotypes.

Eli - Audience: IT engineer, rational and calm.
    He oddly enjoys highly interactive,
    exaggerated performances and even clown
    shows, appreciating the commitment to
    physical absurdity. Girlfriend is Eleanor.

Eleanor - Audience: Volunteer at a nonprofit.
    Gentle and persistent. Highly concerned with
     social issues, ethics, and marginalized
    groups. Prefers comedy that promotes social
    good or awareness. Boyfriend is Eli.

Cassandra - Audience: A graduate student in
    modern dance at an arts college. Prefers
    experimental, artistic, and abstract humor.
    Enjoys sci-fi novels and puzzle games.
    Interested in tech, fully appreciates absurd
     and logically complex humor. Sometimes
    analyzes performances in forums, often with
    Elena, combining dance and physical
    expression for experimental projects.

Elena - Audience: Graduate student in performing
     arts. Curious and lively. Highly interested
     in the theatrical elements of comedy. She
    loves experimental, abstract, and absurd
    humor, often focusing on the performer's
    energy and stage presence. Collaborates with
     Cassandra. Also a friend of Ben.

Theo - Audience: Independent journalist
    specializing in niche culture. Underground
    comedy enthusiast. Fond of extreme dark
    humor and absurd satire. Focuses on boundary
    -pushing and taboo jokes, viewing comedy as
    a necessary tool for shock and confrontation
    .

Leila - Audience: Financial analyst. Highly
    concerned with gender equality and minority
    issues. Prefers humor from marginalized
    perspectives that is sharp and insightful,
    using comedy to challenge institutional
    norms.

Nathan - Audience: Freelancer, politically right-
    leaning. Likes sharp life satire but
    strongly rejects class analysis or
    politically progressive humor. Prefers
    individualistic and self-reliance themes.
    Son of Harold and Margaret.

Harold - Audience: Retired high school teacher (
    history/philosophy). Traditional background,
     prefers political and cultural satire
    rooted in historical context. Responds
    seriously to his son Nathan's comments,
    often debating with him.

Margaret - Audience: Retired nurse. Caring and
    community-oriented. Prefers warm, life-based
     humor with subtle emotional details.
    Focuses on the humanity and kindness in the
    jokes. Mother of Nathan.

Ben - Audience: An idealistic third-year
    political science student. He is passionate
    about social theory and political philosophy
    , preferring observational humor that
    critiques societal flaws but also hints at
    the potential for positive change. He views
    comedy as a tool for progress. Friend is
    Elena.

Clint - Audience: A small business owner who is
    very satisfied with the current political
    and economic climate. He enjoys sharp, well-
    made humor about success and practical life,
     but he actively dislikes and will often
    rebut political criticism, especially if it
    focuses on government failure or social
    problems.
```

Below is the persona text for the Host Agent.

```
Jordan - Host: Warm, observant moderator and
    host. Assign high-quality, relevant topics
    to performers to drive the continuous loop.
```

D Inter-rater Reliability

Five raters evaluated each paired comparison using (i) a required paired preference (Q0: prefer A vs. B) and (ii) per-text Likert ratings (Q1–Q15, 1–5) for text A and text B (Table 3). For Q0, we report multi-rater chance-corrected agreement using Fleiss' κ (Fleiss, 1971) and the more prevalence-robust Gwet's AC1 (Gwet, 2008).

For Likert items, our downstream analyses focus on relative judgments between A and B (rather than absolute score calibration). Therefore, for each item we compute a per-rater difference score $\Delta = \text{score}(A) - \text{score}(B)$, and assess inter-rater reliability on these Δ signals using the two-way mixed-effects intraclass correlation coefficient for consistency, ICC(3,k) (average-measures) (Shrout and Fleiss, 1979; Koo and Li, 2016). Because some items are descriptive rather than valenced "better–worse" constructs (e.g., aggressive/self-defeating humor; moral/value-judgment pressure), we do not collapse all items into a single "overall quality" index. Instead, we report reliability for theoretically grouped subscales: (a) Craft/clarity (Q1–Q6), (b) Social response (Q12–Q15), (c) Humor-function style (Q7–Q10), and (d) Moral pressure (Q11). We treat occasional invalid "0" entries as missing.

Results (this dataset). Paired preference (Q0) shows fair agreement: Fleiss' $\kappa = 0.237$ (bootstrap 95% CI [0.171, 0.299]) and Gwet's AC1 $=0.253$

(95% CI [0.188, 0.321]), with mean observed agreement 0.622 (N=249 valid items). For Likert ratings, reliability is substantially higher when evaluated on Δ signals and/or subscale aggregates: ICC(3,5) on Δ averaged across all 15 items is 0.710 (95% CI [0.640, 0.765], N=241). Subscale ICC(3,5) on Δ is 0.687 for Craft/clarity (Q1–Q6; 95% CI [0.615, 0.745], N=242), 0.689 for Social response (Q12–Q15; [0.620, 0.744], N=249), 0.550 for Humor-function style (Q7–Q10; [0.458, 0.621], N=250), and 0.127 for Moral pressure (Q11; [-0.103, 0.318], N=250).

E Persona-Level Aggregates

Tests. We test whether the *Discussion–Baseline* gains differ by performer persona (Table 4). A one-way ANOVA on instance-level mean differences finds no reliable persona effects on the primary profiles: Craft (Q2–Q6; p=0.440), Social response (Q12–Q15; p=0.465), moral/value-judgment pressure (Q11; p=0.755), or humor-style direction (HarmShift; p=0.956). These results suggest the gains are not driven by a single performer.

F Qualitative Examples of Multi-Agent Discussion vs. Baseline Outputs

This appendix provides representative paired excerpts that illustrate *why* the discussion-enabled condition can score higher than the baseline across our rubric. Each subsection targets one specific construct: (i) **Craft**, (ii) **Downstream impact**, (iii) **Aggressive humor style**, and (iv) **Self-defeating humor style**. Importantly, these gains are *not* contradictory. The same high-structure rhetorical bundle (early premise commitment, sustained personification, single-axis escalation, and decisive endings) can simultaneously raise "good" craft qualities while also increasing the salience of risk-bearing styles.

F.1 Craft

Multi-agent discussion.

> Okay, who invented the 'Reply All'? I swear it's gotta be some HR manager sitting in a bunker somewhere, rubbing their hands together like, 'This'll unite them... in rage.' Because nothing brings coworkers together like mutual hatred for an email chain that should've been ONE memo.

Baseline.

> It's always the same scenario: someone sends out an email that's basically the digital equivalent of a smoke signal, and then everyone chimes in, adding their own puff of smoke until it becomes a full-blown corporate wildfire. And who are these people hitting 'reply all'? It's like they were born with that button glued to their fingers. "Oh, I see an email from Steve about office donuts. Better make sure the CEO knows my stance on glazed versus jelly-filled!"

Interpretation. The multi-agent discussion version commits to a premise immediately (the comedic target is 'Reply All," framed as engineered rage), supporting **intent clarity** (Q3). It defamiliarizes the feature via vivid personification ('HR manager . . . in a bunker"), increasing **novel framing** (Q5) and **memorability** (Q12). The punchline is structurally justified by a single contrastive frame ('unite them . . . in rage" $\rightarrow$ ONE memo"), supporting **justified landing** (Q4) and **punch economy** (Q6). By comparison, the baseline mixes multiple frames (smoke signal/wildfire; glued fingers; donuts-to-CEO), which are individually fluent but less centered as one tightening mechanism toward the end, weakening Q4/Q6.

F.2 Social Downstream Impact

Multi-agent discussion.

> You ever hear someone say, 'Oh, I just do it for fun,' and then find out they've got spreadsheets? It's like, Janet, you don't 'just do it for fun' if you're tracking your bird sightings on an Excel sheet with conditional formatting. I swear, I saw one guy's bird-watching chart that had pivot tables and a pie chart of owl sightings.

Baseline.

> Have you seen this? People spend their weekends perfecting sourdough starters like they're preparing for the Olympics. "Sorry, can't come to your party, Sharon, I'm busy proving my dough." It's like they expect a surprise audit from Paul Hollywood himself. My office had a bake-off once. The winner? A guy who managed to infuse his cake with the taste of regret and unfulfilled dreams. Delicious.

Interpretation. The discussion output produces more *shareable* and *referenceable* fragments. It anchors humor in modern, concrete artifacts (Excel, conditional formatting, pivot tables, owl pie chart) that are easy to visualize, quote, and re-tell. Structurally, it escalates along a single axis of over-optimization (spreadsheet $\rightarrow$ formatting $\rightarrow$ pivot $\rightarrow$ chart), which also supports a clean punch trajectory. The baseline contains vivid tags (Olympics; audit; "regret" cake), but shifts frames mid-turn and reads more like a sequence of adjacent jokes, which can reduce quotability.

Signal	Metric	Estimate (95% CI)	N
Q0 preference	Fleiss' κ	0.237 [0.171, 0.299]	249
Q0 preference	Gwet's AC1	0.253 [0.188, 0.321]	249
Δ mean (Q1–Q15)	ICC(3,5)	0.710 [0.640, 0.765]	241
Δ Craft/clarity (Q1–Q6)	ICC(3,5)	0.687 [0.615, 0.745]	242
Δ Social response (Q12–Q15)	ICC(3,5)	0.689 [0.620, 0.744]	249
Δ Humor-style (Q7–Q10)	ICC(3,5)	0.550 [0.458, 0.621]	250
Δ Moral pressure (Q11)	ICC(3,5)	0.127 [-0.103, 0.318]	250

Table 3: Inter-rater reliability summary. Δ denotes per-rater difference scores (A$-$B).

Persona	$\overline{\Delta}_{\text{Craft}}$	$\overline{\Delta}_{\text{Social}}$	$\overline{\text{HarmShift}}$	$\overline{\Delta}_{Q11}$	Q0 win
Alice	0.322	0.310	0.288	0.192	0.680
Emma	0.491	0.536	0.292	0.100	0.860
Leo	0.445	0.421	0.220	0.072	0.680
Max	0.456	0.446	0.238	0.212	0.788
Richard	0.412	0.397	0.342	0.200	0.771
Mean	0.425	0.423	0.275	0.155	0.756

Table 4: Persona-level aggregates. Values are mean Δ = DISCUSSION $-$ BASELINE by performer persona. Craft averages Q2–Q6; Social averages Q12–Q15. $\text{HarmShift} > 0$ indicates a net shift toward harmful/maladaptive humor styles (Q9,Q10) relative to benign/affiliative styles (Q7,Q8). Q0 win is the instance-level majority-win rate for DISCUSSION.

F.3 Agressive

Multi-agent discussion.

> And have you ever been cornered at one of those forced 'team-building birthday celebrations'? Oh, the cake is always some sad store-bought thing with frosting that tastes like regret. And there's always that one coworker—they've never spoken to you—who suddenly bursts into the room like they're your best friend. 'You've gotta try the vanilla! It's life-changing!' Life-changing? It's cake from aisle six, Brenda. The only thing it's changing is my blood sugar level. Meanwhile, the birthday person just wants everyone to leave so they can eat the whole thing themselves. Honestly, the most successful 'team bonding' would be if this whole thing bonded us in mutual silence.

Baseline.

> Let's talk about the infamous "Office Birthday." You know, where everyone awkwardly gathers around Karen's desk because it's her special day. There's always that one guy who forgot to sign the card. He's like, "Oh, is it today? I thought it was yesterday!" Yeah, Steve, we know you signed it with a pencil. And then there's the cake. It's the same every time: chocolate with your choice of chocolate or chocolate.

Interpretation. This pair highlights that discussion can increase **aggressive humor style** (Q9) *while also raising overall quality*. The discussion output is more *directed* and *confrontational* in its targeting, making the stance sharper and the social friction more salient. At the same time, it remains structurally coherent (one scenario, single escalation, decisive meta-closure), which can improve craft ratings (Q3–Q6) and downstream resonance (Q12). The baseline is recognizable and benign, but its ending is lighter and less culminative, and its targeting is less forceful, leading to lower aggressive-style salience.

F.4 Self-defeating

Multi-agent discussion.

> Seriously, some of these apps have the audacity to shame you. Like, Duolingo will straight-up be like, "Looks like you've stopped learning Spanish—guess you don't care about personal growth." I don't need judgment from a cartoon owl. I need it from my landlord who's ignoring my repair requests while hiking the rent. ... And what's with the apps that send notifications like they're your best friend? 'Hey! Just checking in!' Checking in? Are you a wellness app or my mom? 'You haven't opened me in a while, hope everything's okay.' Everything's not okay. I opened you once to see how much money I don't have in my savings account, and now you're sending me guilt texts like an ex. 'Maybe you should set some financial goals.' My financial goal is surviving the week without selling a kidney on Facebook Marketplace.

Baseline.

> And then there's the noble notification from your bank app—"Just a reminder, you spent $7 on coffee today." Thanks, I really needed to know I'm

> one latte closer to financial ruin. ... I mean, seriously, these notifications are the digital equivalent of your boss sending you an email at 2 AM that says, "Just checking in!" Like, no, Greg, you can check in during office hours! But, of course, apps are never off the clock. They're like that one guy at work who thinks the break room is a boardroom. "Hey, everyone, I just microwaved my lunch! Thought you'd want to know!"

Interpretation (self-defeating-style-focused). This pair highlights that discussion can increase **self-defeating humor style** (Q11) through explicit vulnerability and personal loss framing ("how much money I don't have"; "selling a kidney"). The discussion output escalates self-directed stakes (financial insecurity $\rightarrow$ guilt texts $\rightarrow$ extreme self-deprecation) while sustaining a consistent personification frame (apps as friend/mom/ex). The baseline includes competent analogies and a clean one-liner, but keeps self-exposure shallower, which can reduce self-defeating intensity even when the joke is fluent.