Can LLM-Powered Multi-Agent Systems Augment Human Creativity? Evidence from Brainstorming Tasks

Kazuma Fukumura
Graduate School of Informatics
Kyoto University
Kyoto, Kyoto, Japan
fukumura.kazuma.58p@st.kyoto-u.ac.jp

Takayuki Ito
Graduate School of Informatics
Kyoto University
Kyoto, Kyoto, Japan
ito@i.kyoto-u.ac.jp

Abstract

This paper investigates whether LLM-powered multi-agent systems can effectively augment human creativity in collaborative brainstorming tasks. Traditional brainstorming methods face persistent challenges including evaluation apprehension and free riding, while existing LLM-based discussion systems suffer from premature convergence and homogeneous perspectives that limit creative exploration. To address these dual challenges, we develop a novel multi-agent brainstorming framework that integrates three key innovations: (1) an extended Issue-Based Information System (IBIS) that adds "Theme" nodes and "Idea-to-Idea" transformation paths to prevent premature convergence and enable systematic idea expansion; (2) dynamic role-playing where multiple distinct agent types each dynamically select from topic-specific personas to ensure diverse perspectives while maintaining coherent discussion structure; and (3) evidence-based information suggestion capabilities using web search to stimulate discussion with objective facts rather than subjective criticism. Our experimental evaluation employed a rigorous counter-balanced design comparing human-only, human-agent collaboration, and agent-only conditions across both general topics and domain-specific topics. Results suggest striking domain-dependent effectiveness. For general topics, agent collaboration appears to enhance human creativity, with improvements in originality scores and participants reporting increased agreement that diverse ideas were generated. BERT embedding visualization suggests enhanced semantic dispersion, indicating stronger divergent thinking as agent-assisted posts spread into previously unexplored conceptual territories. The system shows promise in addressing traditional brainstorming challenges: evaluation apprehension appears to decrease as participants report reduced social pressure, and free riding diminishes with the majority of participants selecting agent contributions as the most stimulating ideas. However, for domain-specific topics requiring specialized knowledge, agent effectiveness appears to diminish substantially. Human response rates to agent posts drop noticeably, while creativity metrics suggest decreases rather than improvements. Expert evaluation reveals that a substantial proportion of agent contributions on domain-specific topics are deemed irrelevant by domain knowledge holders, suggesting limitations in current LLM capabilities for specialized contexts where agents lack access to tacit organizational knowledge and domain expertise. This work contributes the first systematic evaluation of structured multi-agent systems for human creativity augmentation, demonstrating both significant potential and important considerations for system design. Our findings provide essential insights for developing effective human-AI collaborative creativity systems: while the combination of IBIS structure, dynamic role-playing, and information suggestion shows promise for general domains, future systems should consider domain-adaptive architectures with specialized knowledge integration to optimize effectiveness across diverse knowledge domains. While our preliminary findings suggest meaningful effects, larger-scale studies will be necessary to establish statistical significance and generalizability across diverse populations and contexts. These results establish initial design principles for next-generation collaborative intelligence systems that can appropriately balance human expertise with AI capabilities.

CCS Concepts

• **Computing methodologies** → **Multi-agent systems**; **Discourse, dialogue and pragmatics**; • **Human-centered computing** → **Collaborative and social computing systems and tools**; **Computer supported cooperative work**.

Keywords

Multi-Agents, Brainstorming, Human-AI Collaboration, Creativity Augmentation, LLM Role-Playing

ACM Reference Format:
Kazuma Fukumura and Takayuki Ito. 2025. Can LLM-Powered Multi-Agent Systems Augment Human Creativity? Evidence from Brainstorming Tasks. In *Collective Intelligence Conference (CI 2025), August 04–06, 2025, San Diego, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3715928.3737479

1 Introduction

Brainstorming aims to generate innovative ideas by aggregating diverse knowledge and perspectives, but faces persistent issues such as "evaluation apprehension," where participants refrain from speaking due to concerns about others' evaluations, and "free riding," where contributions become unbalanced [3].

Meanwhile, while research on LLM applications for discussion support and creativity has been advancing [6], LLM discussions face two major challenges: (1) tendency for premature convergence, and (2) homogeneous perspectives leading to uniform responses [4], [16].

CI 2025, San Diego, CA, USA

ACM ISBN 979-8-4007-1489-4/25/08
https://doi.org/10.1145/3715928.3737479

To address these challenges, we propose a multi-agent approach with three key innovations. First, we extend the Issue-Based Information System (IBIS) framework to structure LLM discussions and prevent premature convergence. Second, we implement dynamic role-playing where agents adopt diverse personas to ensure multifaceted perspectives. Third, we integrate web search capabilities to provide evidence-based information during discussions. Our approach explicitly addresses traditional brainstorming problems of evaluation apprehension and free riding through autonomous agent participation, evaluated via post-experiment questionnaires measuring participants' psychological comfort and engagement.

Realizing this approach is expected to support human creative thinking while expanding possibilities for human-AI collaboration, contributing to qualitative improvement of collective intelligence across diverse fields including educational environments, corporate innovation activities, and research and development.

2 Background

2.1 Brainstorming

Brainstorming is a method proposed by Osborn[15]. To enhance effectiveness, Osborn established four rules:

(1) Generate many ideas
(2) Do not criticize others' ideas
(3) Express ideas freely
(4) Combine and improve ideas

Additionally, Osborn proposed a checklist for developing new ideas from existing ones. This work conducts experiments in accordance with these four rules.

2.2 Issue-Based Information System(IBIS)

The Issue-Based Information System (IBIS) was developed by Kunz and Rittel to structure discussions around wicked problems [1, 10]. The framework represents discourse as a directed graph with four node types: Issues, Ideas, Pros/Cons, and specific relationships between them. Figure 1 illustrates this structure. IBIS has been successfully implemented in tools like gIBIS and Compendium, where the explicit Issue-Position-Argument structure helps teams surface alternative viewpoints and maintain productive exploration of the problem space.

In LLM-based discussions, this structural property is particularly valuable for mitigating LLMs' tendency toward rapid convergence, ensuring sustained exploration of diverse solution paths.

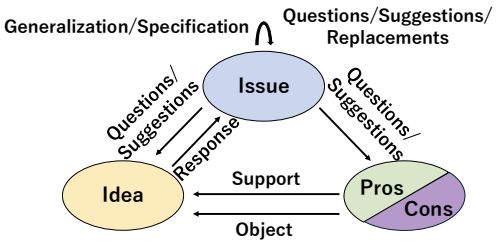

Figure 1: Graph representing IBIS

2.3 LLM role-play

LLMs have the potential to improve performance in various aspects through role-playing behaviors. The ability to imitate specific roles has been shown to induce human-like behavior[17], [19][11] and specifically, the ability to tackle complex tasks[11][2][21].

This work leverages LLM role-playing capabilities to enhance creativity and prevent homogeneous discussions among LLMs.

2.4 Related Work

Lu et al. [12] introduced discussion frameworks and role-playing techniques to enhance LLM creativity. The proposed "LLM Discussion" framework consists of initiation, discussion, and convergence phases. In contrast, this work supports humans through IBIS to address premature LLM discussion convergence while implementing dynamic role-playing.

Wang et al.[20] proposed "Solo Performance Prompting," dynamically simulating multiple personas within a single LLM to improve problem-solving through self-coordination. This work conducts experiments with multi-agents rather than single agents, includes human participants, and aims to support emergent discussions among participants.

Nomura et al. [13] implemented a brainstorming support system utilizing IBIS to address challenges of free riding, social loafing, and social inhibition. In this work, we further extend IBIS and simultaneously incorporate agent role-playing and information suggestions by agents.

Our work distinguishes itself by extending IBIS with "Theme" nodes and "Idea-to-Idea" transformation paths while simultaneously introducing dynamic multi-agent role-playing. This combination addresses both structural and perspective limitations of existing approaches. Unlike single-agent role-playing systems, our multi-agent architecture with six distinct agent types operating along different IBIS paths ensures sustained diversity throughout the discussion process. Furthermore, our integration of evidence-based information suggestion capabilities provides objective fact-based stimulation rather than subjective exchanges, addressing a gap not covered by existing brainstorming support systems.

3 Brainstorming Support Methods Using IBIS and Role-Playing

3.1 Discussion Structure Extending IBIS

We extended the original IBIS structure to better support brainstorming activities. Figure 2 illustrates our modifications, which include three key changes based on preliminary experiments: (1) Addition of a "Theme" node as the root discussion topic; (2) Introduction of "Idea-to-Idea" paths inspired by the SCAMPER method [5] to enable systematic idea expansion through substitution, combination, adaptation, modification, and other creative transformations; (3) Removal of "Issue-to-Pros/Cons" paths as these created redundant exchanges that did not enhance brainstorming productivity.

Our approach employs six distinct agent types, each corresponding to a specific path in the extended IBIS:

- **Idea-to-Theme agent**: Generates initial ideas responding directly to the brainstorming theme

- **Issue-to-Idea agent**: Poses clarifying questions about existing ideas to deepen understanding
- **Idea-to-Issue agent**: Provides answers to questions, elaborating on proposed concepts
- **Idea-to-Idea agent**: Systematically transforms existing ideas using SCAMPER principles (Substitute, Combine, Adapt, Modify, Magnify, Minify, Put to other uses, Eliminate, Reverse, Rearrange)
- **Pros-to-Idea agent**: Provides supporting arguments augmented with web search evidence
- **Cons-to-Idea agent**: Offers constructive counterpoints based on objective web search findings, carefully avoiding subjective criticism to maintain Osborn's brainstorming principle of "no criticism"

Each agent performs role-playing as described in section 3.2.

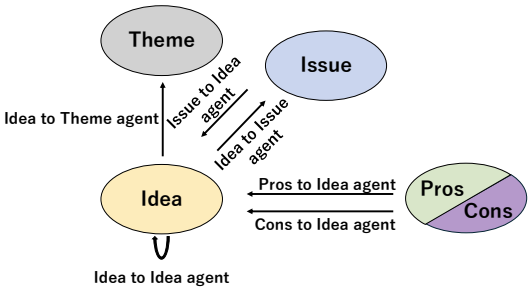

Figure 2: Extended IBIS proposed in this work

3.2 Agent Role-playing and Interaction Logic

Dynamic Persona Selection. As demonstrated by Wang et al., LLMs can achieve enhanced reasoning capabilities through role-playing based on domain-specific dynamic personas. We implement a dynamic persona assignment system where each agent dynamically selects one of 9 topic-specific personas for each response. Preliminary experiments revealed that generating new personas for each response led to excessive diversity that compromised user experience coherence. Therefore, we use a fixed set of 9 diverse personas generated by GPT-4o using few-shot prompting at session initiation. For example, for "Ways to increase visitors to local tourist destinations," generated personas include a Digital Marketer (focused on data-driven customer acquisition), an Environmental Activist (emphasizing sustainable tourism), and a Local Resident Representative (balancing tourism benefits with community impact). This dynamic assignment prevents homogeneous responses while maintaining structural diversity through IBIS roles.

Agent Interaction Logic and Timing. Our approach employs a sophisticated interaction model to ensure balanced participation. Agents respond to human posts according to the following protocol: Idea-to-Theme agent, Idea-to-Issue agent, and Pros-to-Idea agent respond to odd-numbered human posts, while Issue-to-Idea agent, Idea-to-Idea agent, and Cons-to-Idea agent respond to even-numbered human posts. Since our experiment involves groups of 3 humans, the human-to-agent ratio becomes 3:9, ensuring that each persona contributes approximately as much as each human participant.

Contextual Information. Agents receive as context the conversation thread from their target post back through the tree structure to the root "Theme" node, ensuring they understand the specific discussion path that led to the post they are responding to. This focused contextual approach enables coherent and contextually appropriate contributions while maintaining the hierarchical structure of the IBIS-based discussion. The approach maintains conversational coherence while introducing diverse perspectives through the persona-role combination.

3.3 Agent Information Suggestion Functionality

In our agent design, special consideration was given to adhering to the basic brainstorming principle: "Do not criticize others' ideas." Specifically, we implemented the Cons to Idea agent to avoid subjective criticism and point out problems based on objective facts derived from Google search results. Furthermore, based on findings of Kinoshita et al.[9], we also implemented functionality for the Pros to Idea agent to utilize Google search results to promote discussion activation.

In the actual process, we generate queries using GPT-4o from the post content being replied to, and include the top 3 URLs from search results and summaries of each URL in the reply content, using tools such as Google's Custom Search JSON API.

4 Experiment

4.1 Experimental Purpose

This experiment evaluates our multi-agent brainstorming approach across three key conditions: human-only, human-agent collaboration, and agent-only brainstorming. We investigate two primary research questions:

(1) **Agent Effectiveness**: Does agent introduction improve brainstorming outcomes, and how do effects differ between general topics versus domain-specific topics?
(2) **Agent-Topic Compatibility**: How does agent performance vary between general topics and domain-specific topics when operating independently?

For the first question, we examine whether agent introduction enhances brainstorming quality (measured by human post count, originality, elaboration, and flexibility) and addresses traditional brainstorming challenges (evaluation apprehension and free riding, measured through questionnaires). We also analyze how these effects vary by topic type using human response rates to agent posts.

For the second question, we compare agent-only performance between general and domain-specific topics using creativity metrics (originality, elaboration) and conduct human expert evaluation to assess whether agent outputs provide meaningful value to domain knowledge holders.

4.2 Experimental Setup

Participants and Recruitment. We recruited university students from Kyoto University as participants. All participants were native Japanese speakers from the Ito Laboratory, possessing domain expertise for domain-specific topics.

Experimental Design. The study used a between-subjects design with within-subject condition comparison. Participants were assigned to five groups of three, with three groups assigned to general topics and two groups assigned to domain-specific topics. This uneven distribution was due to practical constraints in recruiting domain experts for specialized topics. Each group completed two 15-minute brainstorming sessions for their assigned topic type:

- Without agents condition: Human participants only
- With agents condition: Human participants + LLM agents

The session order was counterbalanced using an AB / BA design to control for learning effects, with a 5-minute break between sessions. Agent-only conditions were conducted separately for general and domain-specific topics to evaluate agent performance without human participation.

Communication Platform. All sessions used the D-Agree platform [7, 8], a web-based discussion system that structures posts in tree format resembling mind maps. D-Agree's asynchronous nature directly addresses "production blocking" (participants waiting for turns to speak) by allowing simultaneous posting. The platform automatically timestamps all contributions and maintains discussion thread relationships for analysis.

General topics. General topics ("Ways to increase visitors to local tourist destinations" and "Ways to increase visitors to theme parks") were selected as domains accessible to both human participants and LLMs through general world knowledge. These topics require creative thinking and diverse perspectives but do not demand specialized domain expertise, making them suitable for evaluating agent effectiveness in broadly accessible knowledge domains.

Domain-specific topics. Domain-specific topics ("Improving Kyoto University Ito Laboratory environment" and "Ito Laboratory selling points") were chosen because all participants were current laboratory members with intimate knowledge of internal operations, research projects, and daily challenges—knowledge unavailable to LLMs pre-trained on public data.

Questionnaire. After each general topic session, participants completed a 15-item questionnaires using 5-point Likert scales and open-ended questions. Evaluation apprehension was evaluated through questions such as "Were there situations where you hesitated to speak?" and "Specifically, in what situations did you hesitate to speak?" Free riding was evaluated using items including "Were you stimulated by statements from others?", "Were there statements from others that you thought were particularly good?", "Which statement from others did you think was the best?" and "Do you think diverse ideas were came up in this brainstorming session?" The complete 15 questionnaire items are provided in the AppendixB.

4.3 Evaluation Criteria

We evaluate brainstorming outcomes using both quantitative and qualitative metrics derived from discussion logs and post-experiment questionnaires.

Quantitative Metrics. To assess the volume and interaction patterns of human participation, we employ two primary quantitative measures.

- **Human Post Count**: Total number of human contributions, excluding agent posts since GPT-4o can generate unlimited responses[14]
- **Human Reply Rate to Agents**: Proportion of human posts that directly respond to agent contributions, calculated as (direct human replies to agent posts)/(total human posts)

Qualitative Metrics from TTCT Framework. For creativity assessment, we employ the Torrance Tests of Creative Thinking (TTCT) framework [18], a widely-established psychological instrument for measuring divergent thinking and creativity. TTCT evaluates creative thinking through four dimensions: fluency (quantity of ideas), flexibility (diversity of categories), originality (novelty and uniqueness), and elaboration (level of detail and development). From this framework, we adapt three metrics relevant to brainstorming evaluation:

- **Originality**: Novelty and creativity of ideas while maintaining relevance (1-5 Likert scale)
- **Elaboration**: Level of detail and depth in idea descriptions (1-5 Likert scale)
- **Flexibility**: Semantic diversity of ideas, measured through BERT embedding visualization using t-SNE and principal component analysis

We excluded fluency from TTCT as it corresponds to post quantity, already captured by our post count metric.

Evaluation Implementation. Originality and elaboration are assessed using GPT-4o with standardized 5-point rubrics, though we acknowledge the potential limitation of using LLM-based evaluation for LLM-generated content. For example, originality score 5 indicates "Extremely original - very unique and rare ideas showing high novelty, creativity, and unexpected elements rarely conceived in typical contexts." For flexibility, we generate BERT embeddings for all posts and visualize semantic dispersion through dimensionality reduction, allowing assessment of idea diversity across the brainstorming space.

Questionnaire Assessment. Post-experiment questionnaires measure evaluation apprehension through questions like "Were there situations where you hesitated to speak?" and free riding through items including "Were you stimulated by others' statements?" and "Do you think diverse ideas were generated?" (complete questionnaire in Appendix B).

5 Results and Discussion

5.1 Overview of Results

Our experimental evaluation reveals differential effectiveness of the multi-agent system across topic types. For general topics, agent introduction appears to enhance human creativity while addressing

traditional brainstorming challenges. However, for domain-specific topics, agent effectiveness diminishes substantially, with agents often generating content perceived as irrelevant by domain experts.

Importantly, across both topic types, agent introduction did not significantly change human post volume, suggesting that benefits arise from qualitative rather than quantitative improvements in human contributions.

5.2 General Topics: System Effectiveness Confirmed

Creativity Enhancement Through Agent Collaboration. For general topics, our approach demonstrates shows promising trends across multiple creativity dimensions based on the TTCT framework. Our analysis reveals improvements in originality, maintained elaboration quality, and enhanced flexibility through semantic diversity expansion.

Originality and Elaboration: Table 1 shows that human posts in agent-assisted conditions achieved substantially higher originality scores (2.47 ± 0.17) compared to human-only conditions (2.13 ± 0.19), representing an average improvement of 16.0%. While elaboration scores showed only modest improvements (from 2.03 ± 0.17 to 2.07 ± 0.13), the primary benefit appears in enhanced creative thinking rather than descriptive detail.

Table 1: Originality and Elaboration: General Topics

Condition	Originality	Elaboration
Without agents	2.13 ± 0.19	2.03 ± 0.17
With agents	2.47 ± 0.17	2.07 ± 0.13

Flexibility and Semantic Diversity: Our flexibility analysis provides compelling visual evidence of enhanced divergent thinking through t-SNE and principal component analysis (PCA). Figure 3 presents these results with blue representing human-only posts, red representing all posts from agent-assisted conditions, and green representing only human posts from agent-assisted sessions.

The t-SNE visualization (left panel) reveals that human posts generated with agents present (green) are distributed more widely than human-only posts (blue), with green points spreading outward from the central cluster into previously unexplored semantic regions. This pattern suggests that agent interaction expanded human thinking into more diverse conceptual territories.

The PCA visualization (right panel) corroborates these findings, showing that agent-introduced conditions (red and green) demonstrate markedly wider distribution than human-only posts (blue). The agent-introduced condition covers the human-only conceptual territory while extending substantially beyond these boundaries, demonstrating that agents enhance brainstorming divergence through increased semantic diversity.

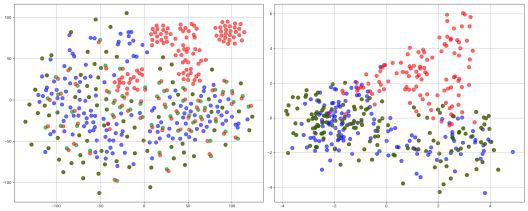

Figure 3: Scatter plot of posts for general topics (Left: t-SNE visualization, Right: PCA visualization)

Addressing Traditional Brainstorming Challenges. Beyond creativity enhancement, our approach successfully addresses key psychological barriers that traditionally limit brainstorming effectiveness. Questionnaire results reveal significant improvements in evaluation apprehension and free riding behaviors.

Evaluation Apprehension Reduction: Agent introduction appears to have decreased participants' hesitation to contribute ideas. Figure 4 shows substantial improvement, with participants reporting reduced social pressure.

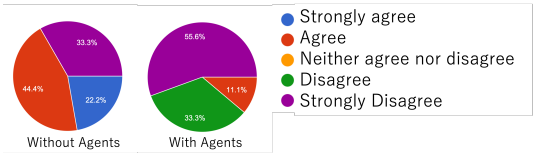

Figure 4: Were there situations where you hesitated to speak?

In response to "When specifically did you hesitate to speak?", multiple participants responded with content such as "I hesitated to speak when new ideas had been exhausted," including opinions such as "I got caught up thinking about others' ideas" and "I got stuck on what to say." In the condition with agents, responses such as "I didn't hesitate to speak because I knew the other was an agent" were observed, suggesting that evaluation apprehension was resolved as expected, though some noted "There were so many agent statements that it took time just to read them, and I didn't have time to speak myself." While complete elimination of evaluation apprehension remains challenging, agents can reduce psychological inhibition in many cases, though adjustment of agent speech amount and timing may be necessary.

Free Riding Mitigation: Three indicators demonstrate substantial improvement in participant engagement. Participants reported increased stimulation from others' contributions (Figure 5), identified more high-quality ideas from others (Figure 6), and most importantly, the perception of idea diversity showed marked improvement (Figure 7). Notably, when asked "Which statement from others did you think was the best?", in 8 out of 9 cases, participants selected agent posts as the best contributions, indicating that agent content successfully stimulated human thinking and promoted new idea expression.

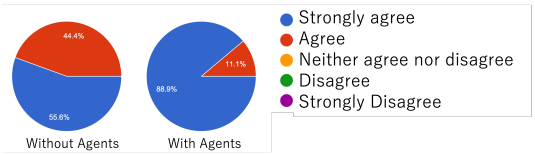

Figure 5: Were you stimulated by statements from others?

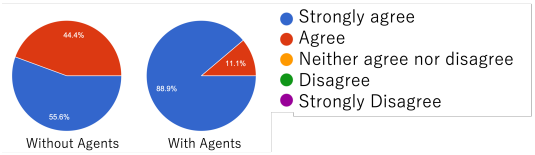

Figure 6: Were there statements from others that you thought were good?

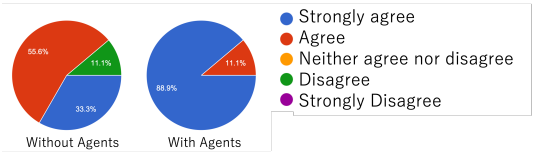

Figure 7: Do you think diverse ideas came up in this brainstorming session?

System Impact Summary. These findings suggest that IBIS structure, role-playing diversity, and information suggestion capabilities may effectively enhance brainstorming for general topics, though larger-scale studies are needed to confirm these preliminary results. Regarding human post frequency, agent introduction showed mixed effects across the three general topic groups: two groups showed decreases of 25.0% and 9.2% respectively, while one group showed an 18.8% increase, indicating no consistent directional change in posting behavior. This suggests that benefits arise from qualitative rather than quantitative improvements in human contributions.

5.3 Domain-Specific Topics: Limitations Revealed

Decreased Creativity Metrics. For domain-specific topics, agent introduction shows markedly different results. Table 2 reveals that agent introduction actually decreased human originality scores (from 2.12 ± 0.08 to 1.82 ± 0.22), contrasting sharply with general topic improvements. Elaboration scores also showed slight decreases (from 1.95 ± 0.09 to 1.85 ± 0.16), indicating no compensatory benefits in idea development.

Similarly, human post frequency for domain-specific topics showed mixed patterns: one group decreased by 23.1% while another increased by 26.6%, again indicating no consistent directional change in posting behavior despite the substantial magnitude of individual group variations.

Table 2: Originality and Elaboration: Domain-specific Topics

Condition	Originality	Elaboration
Without agents	2.12 ± 0.08	1.95 ± 0.09
With agents	1.82 ± 0.22	1.85 ± 0.16

Most concerning, Figure 8 demonstrates that agent-assisted conditions fail to capture the full conceptual space of human-only brainstorming. The visualization reveals that agent conditions (red, green) miss significant portions of the human-only distribution (blue), suggesting that agent participation actually constrains rather than expands human thinking in domain-specific topics.

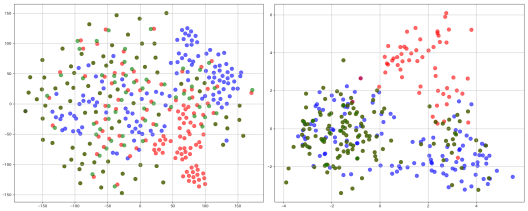

Figure 8: Scatter plot of posts for domain-specific topics (Left: t-SNE visualization, Right: PCA visualization)

Reduced Human-Agent Interaction. Human engagement with agent content drops substantially for domain-specific topics. Response rates to agent posts decrease from 0.48±0.19 (general topics) to 0.21 ± 0.13 (domain-specific topics). This dramatic reduction indicates that participants with domain expertise recognize agent contributions as less relevant or valuable, leading to decreased interaction and potential inhibition of human creativity.

Underlying Causes. Three factors explain these domain-specific limitations:

- **Knowledge Gaps**: Agents lack access to organization-specific information (e.g., laboratory operations) absent from training data
- **Contextual Understanding**: Limited context from discussion threads cannot capture tacit knowledge and shared understanding
- **Role-Playing Mismatch**: Generated personas may lack sufficient specialization for domain-specific topics

5.4 Agent-Only Performance: Domain-Specific Knowledge Challenges

Independent agent evaluation confirms the domain-specific knowledge limitations observed in human-agent collaboration. When comparing agent-only brainstorming across topic types, domain-specific topics consistently underperform general topics in creativity metrics.

Statistical analysis reveals significant differences in agent creativity across domains. Mann-Whitney U testing shows significantly lower originality scores for domain-specific topics ($p=0.005196$), while elaboration scores show no significant variation ($p=0.8534$).

This pattern suggests that while agents maintain descriptive capabilities across domains, their creative insight diminishes substantially when domain-specific knowledge is required.

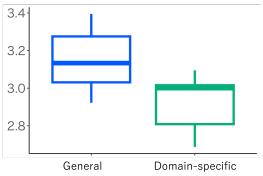

Figure 9: Box plot of originality

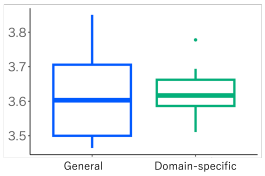

Figure 10: Box plot of elaboration

Human expert evaluation provides crucial validation of these automated metrics. Domain experts (n=4) assessed randomly selected agent-only logs from domain-specific topic sessions, evaluating each post for relevance and meaningfulness within the knowledge domain. Results show that over 50% of agent posts received irrelevant ratings from the majority of evaluators, with 43-40% deemed irrelevant by 3+ evaluators and 7-35% considered irrelevant by all 4 evaluators (Table 3).

Table 3: Evaluation of the proportion of off-target posts by agents

Number of people who answered "yes"	Proportion in Sample 1	Proportion in Sample 2
3 people	43%	40%
4 people	7%	35%

These findings indicate fundamental limitations in current LLM capabilities for specialized domain applications, where agents lack access to tacit knowledge, organizational context, and domain-specific expertise required for meaningful contribution.

5.5 Implications and Future Directions

System Design Insights. Our findings reveal that the effectiveness of LLM-based brainstorming support systems is highly domain-dependent. The combination of IBIS structure, dynamic role-playing, and information suggestion proves highly effective for general topics, where diverse perspectives and external information can stimulate creative thinking. However, the same mechanisms become counterproductive in specialized domains where agents lack requisite knowledge depth.

Mechanisms of Success and Failure. For general topics, success stems from three complementary mechanisms:

(1) **Perspective Diversification**: Dynamic role-playing introduces viewpoints participants might not naturally consider
(2) **Information Augmentation**: Web search capabilities provide evidence and inspiration beyond participants' immediate knowledge
(3) **Social Inhibition Reduction**: Agent presence reduces evaluation apprehension while maintaining engagement

For specialized domains, these same mechanisms create obstacles:

(1) **Irrelevant Perspectives**: Role-playing generates personas lacking domain expertise
(2) **Superficial Information**: Web search yields general information rather than domain-specific insights
(3) **Expert Frustration**: Domain experts recognize agent limitations, leading to disengagement

These findings contribute essential insights for developing effective human-AI collaborative creativity systems, demonstrating both the significant potential and critical limitations that must be addressed through domain-specific system design.

6 Limitations

Our study has several important limitations that must be acknowledged when interpreting these findings.

Sample Size and Statistical Power. With only 5 groups, our human-involved experiments have limited statistical power for definitive significance testing. The small sample size constrains the generalizability of our findings and prevents robust statistical analysis of interaction effects between conditions. Future work will conduct larger-scale studies with multi-site recruitment to achieve adequate statistical power for comprehensive effect size analysis.

Domain Knowledge Scope. Our "domain-specific knowledge" evaluation was restricted to laboratory-specific topics familiar only to participants. This narrow domain definition limits our understanding of how our approach might perform across broader professional domains such as medicine, law, or engineering, where domain-specific knowledge requirements differ substantially. We plan to develop domain-specific agent training protocols and validate our approach across diverse professional contexts with domain experts.

Temporal Constraints. The 15-minute session duration may not capture longer-term creativity dynamics or system sustainability. Real-world brainstorming often requires extended engagement, and our brief sessions cannot assess whether the observed benefits persist over longer durations or whether participant fatigue affects human-agent interaction quality. Future work will conduct longitudinal studies with multi-session experiments to examine temporal dynamics and develop adaptive intervention strategies.

Evaluation Methodology Constraints. Our creativity assessment relies heavily on automated GPT-4o evaluation, which may introduce systematic biases. While we implemented inter-rater reliability measures, the lack of human expert evaluation for all conditions limits the validity of our creativity metrics, particularly for domain-specific topics. Future research will develop hybrid evaluation frameworks combining automated assessment with diverse human expert judgments and establish domain-specific creativity benchmarks.

System Architecture and Scalability. Our current implementation uses a fixed set of 9 personas and 6 agent types, which may not scale effectively to different group sizes or brainstorming contexts. The observed challenges with information overload (participants

reporting difficulty processing agent content) suggest that agent participation strategies need refinement. Future implementations should explore adaptive participation mechanisms that adjust agent involvement based on real-time assessment of human engagement and cognitive load.

These limitations collectively point to the need for more sophisticated, domain-aware systems that can dynamically adapt their support strategies while maintaining rigorous evaluation standards that do not rely solely on the same technologies being evaluated.

7 Conclusion

This work addresses fundamental challenges in human-AI collaborative creativity by developing a novel multi-agent brainstorming support system that integrates extended IBIS structure, dynamic role-playing, and information suggestion capabilities. Our experimental evaluation reveals both significant potential and critical limitations for LLM-powered collaborative creativity systems.

We demonstrated preliminary evidence that multi-agent systems can enhance human creativity in general domain brainstorming through perspective diversification via dynamic role-playing, information augmentation through web search, and social disinhibition where agent presence reduces evaluation apprehension. For general topics, our approach showed promising improvements in human post originality, semantic diversity, and participant perception of idea diversity, while showing potential for addressing traditional brainstorming challenges.

However, our findings reveal a critical limitation: system effectiveness is highly domain-dependent. For domain-specific topics, the same mechanisms that enhance general topic creativity become counterproductive. Agent introduction decreased human creativity metrics, and expert evaluation revealed that many agent contributions were deemed irrelevant by domain experts. This contrast illuminates fundamental challenges in current LLM capabilities for domain-specific applications, where agents lack access to tacit knowledge and domain expertise.

Our work provides crucial insights for developing effective human-AI collaborative creativity systems. While appropriately designed multi-agent systems can overcome traditional brainstorming limitations for general domains, the domain-specific limitations highlight the need for domain-adaptive architectures.

The significant contribution of this work lies in demonstrating measurable creativity enhancement for general domains while revealing the boundaries of current LLM capabilities. As human-AI collaboration becomes increasingly prevalent across diverse professional contexts, understanding these domain-dependent effectiveness patterns is essential for developing systems that truly augment rather than constrain human creative potential.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP22H00533, JST Strategic International Collaborative Research Program (SICORP) JPMJSC2307, and JST CREST JPMJCR20D1.

References

[1] Jeff Conklin. 2003. Dialog mapping: Reflections on an industrial strength case study. In *Visualizing Argumentation*. Springer London, London, 117–136.

[2] Christopher Cui, Xiangyu Peng, and Mark Riedl. 2023. Thespian: Multi-character text role-playing game agents. *arXiv [cs.AI]* (Aug. 2023).

[3] Michael Diehl and Wolfgang Stroebe. 1987. Productivity loss in brainstorming groups: Toward the solution of a riddle. *J. Pers. Soc. Psychol.* 53, 3 (Sept. 1987), 497–509.

[4] Ning Ding, Yulin Chen, Bokai Xu, Yujia Qin, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. 2023. Enhancing Chat Language Models by Scaling High-quality Instructional Conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 3029–3051. doi:10.18653/v1/2023.emnlp-main.183

[5] Bob Eberle. 1996. *Scamper on: Games for imagination development*. Prufrock Press Inc.

[6] Carlos Gómez-Rodríguez and Paul Williams. 2023. A Confederacy of models: A comprehensive evaluation of LLMs on creative writing. *arXiv [cs.CL]* (Oct. 2023).

[7] Takayuki Ito, Rafik Hadfi, and Shota Suzuki. 2022. An agent that facilitates crowd discussion: A crowd discussion support system based on an automated facilitation agent. *Group Decis. Negot.* 31, 3 (June 2022), 621–647.

[8] Takayuki Ito, Shota Suzuki, Naoko Yamaguchi, Tomohiro Nishida, Kentaro Hiraishi, and Kai Yoshino. 2020. D-Agree: Crowd Discussion Support System Based on Automated Facilitation Agent. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 09 (Apr. 2020), 13614–13615. doi:10.1609/aaai.v34i09.7094

[9] Ryosuke Kinoshita and Shun Shiramatsu. 2022. Agent for recommending information relevant to web-based discussion by generating query terms using GPT-3. In *2022 IEEE International Conference on Agents (ICA)*. IEEE, 24–29.

[10] Werner Kunz and Horst WJ Rittel. 1970. *Issues as elements of information systems*. Vol. 131. Citeseer.

[11] Cheng Li, Ziang Leng, Chenxi Yan, Junyi Shen, Hao Wang, Weishi Mi, Yaying Fei, Xiaoyang Feng, Song Yan, Haosheng Wang, Linkang Zhan, Yaokai Jia, Pingyu Wu, and Haozhen Sun. 2023. ChatHaruhi: Reviving Anime character in reality via large language model. *arXiv [cs.CL]* (Aug. 2023).

[12] Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung yi Lee, and Shao-Hua Sun. 2024. LLM Discussion: Enhancing the Creativity of Large Language Models via Discussion Framework and Role-Play. In *First Conference on Language Modeling*. https://openreview.net/forum?id=ybaK4asBT2

[13] Moeka Nomura, Takayuki Ito, and Shiyao Ding. 2024. Towards collaborative brain-storming among humans and AI agents: An implementation of the IBIS-based brainstorming support system with multiple AI agents. In *Proceedings of the ACM Collective Intelligence Conference*, Vol. 24. ACM, New York, NY, USA, 1–9.

[14] OpenAI. 2024. GPT-4o System Card. *arXiv [cs.CL]* (Oct. 2024).

[15] Alex F Osborn. 1953. Applied imagination. *Charles Scribner* (1953).

[16] Vishakh Padmakumar and He He. 2024. Does Writing with Language Models Reduce Content Diversity?. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Feiz5HtCD0

[17] 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*. ACM, New York, NY, USA.

[18] Ellis Paul Torrance. 1966. *Torrance Tests of Creative Thinking. Norms-Technical Manual. Research Edition. Verbal Tests Forms a and B. Figural Tests Forms a and B.* Personnel Press.

[19] Xintao Wang, Yunze Xiao, Jen-Tse Huang, Siyu Yuan, Rui Xu, Haoran Guo, Quan Tu, Yaying Fei, Ziang Leng, Wei Wang, Jiangjie Chen, Cheng Li, and Yanghua Xiao. 2024. InCharacter: Evaluating personality fidelity in role-playing agents through psychological interviews. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 1840–1873.

[20] Zhenhailong Wang, Shaoguang Mao, Wenshan Wu, Tao Ge, Furu Wei, and Heng Ji. 2024. Unleashing the Emergent Cognitive Synergy in Large Language Models: A Task-Solving Agent through Multi-Persona Self-Collaboration. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 257–279. doi:10.18653/v1/2024.naacl-long.15

[21] Zekun Moore Wang, Zhongyuan 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, Stephen W Huang, Jie Fu, and Junran Peng. 2023. RoleLLM: Benchmarking, eliciting, and enhancing role-playing abilities of Large Language Models. *arXiv [cs.CL]* (Oct. 2023).

A Interface and Interaction Examples

A.1 D-Agree Platform Interface

Figure 11 was *not* part of the experimental log; it is an example thread that we crafted to familiarize participants with the interface before the study began. Because every icon in the left margin represents a specific IBIS node, participants can immediately recognize whether a post plays the role of an *Issue*, proposes a new *Idea*, or supplies *Pros/Cons*. This visual mapping, together with the dotted response lines that preserve the parent–child hierarchy, enables users to follow the discussion while simultaneously internalizing the IBIS structure, an ability that proved helpful when they later engaged in the real brainstorming tasks in D-Agree.

Figure 11: D-Agree platform interface showing IBIS structure with icons and reply connectors

A.2 Sample Human-Agent Interaction

The following transcript excerpt illustrates typical human-agent interaction dynamics during our brainstorming experiment for the general topic "Ways to increase visitors to theme parks." This example demonstrates how the extended IBIS structure facilitates structured dialogue while dynamic role-playing introduces diverse perspectives to stimulate creative thinking.

Figure 12: Actual experiment log showing human-agent interaction in D-Agree platform.

The interaction sequence below showcases several key features of our system:

```
[AGENT - Issue]: As a regional event producer, I have a
question: How will you decide on the target audience for the
 commercial, and do you have any ideas for incorporating
local characteristics into its content?

[HUMAN - Idea]: One idea is to target the commercial at
students and families with children who are likely to enjoy
the theme park. To incorporate local character, we could
highlight the park's unique features.

[AGENT - Issue]: As a first-time visitor, I wonder whether
the commercial will make it clear that newcomers can enjoy
it easily. Specifically, what kind of content are you
planning to include?
```

This exchange exemplifies several important aspects of our system design:

Role-Playing Effectiveness: The first agent adopts the persona of a "regional event producer," bringing domain-relevant expertise while asking targeted questions that guide discussion depth. The second agent takes the perspective of a "first-time visitor," introducing a different stakeholder viewpoint that challenges participants to consider accessibility and user experience.

IBIS Structure Implementation: Both agent contributions follow the Issue-to-Idea path in our extended IBIS framework, posing clarifying questions that deepen understanding of the human participant's initial idea. This structured approach prevents premature convergence while encouraging systematic exploration of concepts.

Discussion Stimulation: The agents' questions prompt the human participant to elaborate on targeting strategies and content

specifics, demonstrating how agents can stimulate more detailed thinking without directly providing solutions. This aligns with brainstorming principles of building upon ideas rather than immediately evaluating them.

Contextual Coherence: Each agent contribution directly relates to the preceding human post while introducing new angles for consideration. The "regional event producer" focuses on strategic marketing aspects, while the "first-time visitor" emphasizes user experience, showing how different personas naturally contribute complementary perspectives.

B Questionnaire items

The content of each questionnaire item is as follows:

(1) Are you satisfied with this brainstorming session? (Common, 5-point scale, required)
(2) Do you think diverse ideas came up in this brainstorming session? (Common, 5-point scale, required)
(3) Do you think you were able to come up with many ideas in this brainstorming session? (Common, 5-point scale, required)
(4) Do you think you were able to come up with creative ideas in this brainstorming session? (Common, 5-point scale, required)
(5) Were there situations where you hesitated to speak? (Common, 5-point scale, required)
(6) Specifically, in what situations did you hesitate to speak? (Common, open-ended, optional)
(7) Were you stimulated by statements from others? (Common, 5-point scale, required)
(8) Were there statements from others that you thought were good? (Common, 5-point scale, required)
(9) Which statement from others did you think was the best? (copy of the statement) (Common, open-ended, required)
(10) Which of your own statements did you think was the best? (copy of the statement) (Common, open-ended, required)
(11) Regarding the statements you mentioned in the previous questions (your own/others'), which statement do you think was the best among all statements? (Common, multiple choice, required)
(12) Do you think agents are beneficial for brainstorming? (Agent condition only, 5-point scale, required)
(13) What were the good points of the system? (Agent condition only, open-ended, optional)
(14) What aspects of the system should be improved? (Agent condition only, open-ended, optional)
(15) Anything else you would like to share? (Common, open-ended, optional)