Creativity Is Not an Outcome: A Process-Based Framework for Mapping Individual Creative Contributions in Human–LLM Collaboration

Nishthaa Lekhi[a], Trevor Patten[b], Prakash Patil[a], Mengyao Li[b] and Areen Alsaid[a,*]

[a]University of Michigan-Dearborn, Dearborn, MI, USA
[b]Georgia Institute of Technology, Atlanta, GA, USA

ARTICLE INFO

Keywords:
Human–AI collaboration
Interactional Creativity
Human Experience
Semantic Divergence
Emotional Change

ABSTRACT

If large language models (LLMs) are reshaping creative contributions in collaboration, the dominance of outcome-based creativity evaluation may obscure these changes by capturing only the final product. This makes it hard to determine whether LLMs support or displace human creative effort. To address this, we introduce a computational framework for tracking human creative contributions during live brainstorming dialogues. Using Observable Creative Sense-Making (OCSM), we operationalize participation, novelty, and appropriateness via computational semantic and linguistic measures evaluated against expert ratings. These enable turn-by-turn analysis of how contributions evolve, offering granularity that static benchmarks lack. In a between-subjects design, we compare creative behavior, subjective experience, and emotional trajectories of participants brainstorming with a human or an LLM. Results show LLM collaboration leads to greater idea divergence and exploration but lower participation and appropriateness. Human–human collaboration demonstrates stronger participation, contextual grounding, and cumulative idea development. Participants collaborating with an LLM also reported less positive experiences and lower confidence in collaborative outcomes. These findings reveal that different partners support different creativity dimensions - tradeoffs invisible to outcome-based assessments. By showing how creative contributions emerge and evolve, our framework helps identify where human creative effort is amplified, redistributed, or diminished. This visibility enables the design of conversational AI systems that strengthen human creative engagement instead of shifting effort away from human collaborators. More broadly, our framework provides a foundation for benchmarking creative contributions, identifying gaps in human–AI collaboration, and evaluating how AI technologies shape human creative potential over time.

1. Introduction

Creativity is more than an outcome. It is the process by which the mind synthesizes experiences, connects disparate ideas, and navigates constraints to produce something that did not exist before. This capacity to recombine the familiar into the unfamiliar, identify structure within ambiguity, and generate new possibilities has long been regarded as a defining aspect of human cognition (Higgins, 1994; Wang, 2009). It is what enabled the development of language, architecture, medicine, and, ultimately, the computational systems now embedded in everyday creative work. Large language models (LLMs), among the most advanced of these systems, are themselves products of human creativity. Yet their increasing integration into creative workflows raises a question that is easy to dramatize but difficult to answer precisely (Franceschelli and Musolesi, 2025; Kim, 2024; Zhou et al., 2024): *Does collaborating with an LLM erode the very creative capacities that produced it?* Framed as an existential threat, this question invites fear-mongering. Framed as a design problem, it becomes tractable. The challenge is not to resist LLMs or to celebrate them, but to understand how their presence in collaborative settings reshapes the way humans contribute creatively, and to use that understanding to amplify and preserve, rather than passively surrender human creative capability. This requires moving past a binary debate over whether LLMs "are" creative and toward a more precise investigation of what happens to human creativity when LLMs are involved.

*Corresponding author.
nlekhi@umich.edu (N. Lekhi); tpatten7@gatech.edu (T. Patten); patilpb@umich.edu (P. Patil); mengyao.li@gatech.edu (M. Li); alsaid@umich.edu (A. Alsaid)
ORCID(s): 0009-0007-7982-0640 (N. Lekhi); 0009-0008-3581-9434 (T. Patten); 0009-0009-3736-6229 (P. Patil); 0000-0002-0819-4693 (M. Li); 0000-0003-2852-9750 (A. Alsaid)

Pursuing this investigation is harder than it might seem. Creativity has traditionally been measured as an outcome: how many ideas can a person generate in a fixed period, how original are those ideas, how useful (Kaufman and Sternberg, 2010; Barron and Harrington, 1981; Torrance, 1974). These metrics are countable and concrete and therefore serve well in controlled laboratory settings built around tasks like the Alternate Uses Task (AUT) or the Divergent Association Task (DAT). They also continue to anchor much of the work comparing human and LLM creative performance (Hubert et al., 2024; Kumar et al., 2025; Vaccaro et al., 2024; Bangerl et al., 2025). By these measures, LLMs often perform impressively, generating large numbers of ideas that meet conventional thresholds of originality and diversity (Kumar et al., 2025; Hubert et al., 2024). However, these outcome-based evaluations reveal little about how creativity emerges during collaboration. This conflation of treating output metrics as evidence of creative capacity obscures more fundamental concerns. First, outcome-based measures cannot capture how contributions develop, shift, or accumulate across an interaction. Second, even within outcome-based studies, aggregate creativity scores collapse the temporal structure of collaboration into a single endpoint evaluation, obscuring the sequential dynamics through which ideas emerge and evolve. Recent work in computational creativity has similarly shown that models demonstrating comparable task performance may nevertheless differ substantially in creative behavior and problem-solving strategies (Kim et al., 2025). This suggests that outcome-based evaluations alone may fail to capture meaningful differences in how creativity emerges throughout a task. Third, there is currently no established framework for tracking creativity at the level of individual contributions throughout human-LLM dialogue while simultaneously examining how these contribution patterns relate to subjective experience and affect. Together, these limitations suggest the need to move beyond evaluating creativity as a final outcome and toward understanding it as a process that unfolds throughout interaction.

Contemporary creativity research emphasizes that creativity unfolds through interaction rather than residing solely in individual cognition or isolated outputs (Sawyer, 2017; Glaveanu et al., 2013; Smuts, 1992). These perspectives have informed frameworks such as Creative Sense-Making (CSM) and Observable Creative Sense-Making (OCSM), which provide methods for evaluating creative contribution throughout interaction by examining creativity through coding cognitive and observable interactional dynamics (Davis et al., 2017; Deshpande et al., 2023, 2024). However, their applications have been limited to embodied collaborative contexts, such as pretend play and collaborative drawing, but their use to text-based Human–LLM dialogue remains underexplored.

Despite growing interest in human-LLM creative collaboration, no study has examined how partnering with an LLM reshapes the participation, novelty, and appropriateness of human contributions across the turns of a live brainstorming dialogue, nor whether those shifts correspond to changes in subjective experience, perceived outcome confidence, or emotional affect. This gap is consequential: prior work raises concerns about diminished autonomy, reduced sense of control, and ambiguity around the experience of AI-supported work (Baldeo, 2026; Biermann et al., 2022). Yet without a process-level account of how human contribution unfolds turn by turn, and how it relates to what the human experiences and feels, these concerns remain difficult to locate, measure, or address in system design. A framework that tracks creativity as it evolves within interaction, rather than evaluating it only at the endpoint, is therefore needed. To address this gap, we introduce a scalable process-based framework for evaluating creativity at the level of individual contributions within text-based collaboration. Building on OCSM, the framework operationalizes participation, novelty, and appropriateness using computational semantic and linguistic measures and evaluates their alignment with expert human ratings. Using this framework, we compare human–human and human-LLM brainstorming interactions to investigate how creative contributions emerge, evolve, and accumulate across different collaborative partners. Beyond evaluating collaborative outcomes, this approach enables a deeper understanding of how LLMs reshape human creative participation and provides a foundation for designing conversational systems that better support human creative potential.

2. Related Work

2.1. Creativity as an Interactional and Collaborative Process

Going beyond traditional views that define creativity primarily through the production of original and useful outcomes (Kaufman and Sternberg, 2010; Barron and Harrington, 1981; Torrance, 1974), and building on interactional perspectives that view creativity as socially situated and emergent through collaboration (Smuts, 1992; Glaveanu et al., 2013), Sawyer introduces the idea of "group genius," where novelty emerges through processes such as turn-taking, uptake, responsiveness, adaptation, reframing, and the continuous transformation of prior contributions during interaction (2017). Similar perspectives across collaborative creativity research further emphasize that ideation

develops recursively through cycles of elaboration, negotiation, mutual influence, and reflection between collaborators (Rouse, 2020; Bryan-Kinns et al., 2007; Lu et al., 2019).

In the context of conversational and text-based collaboration, these interactional processes similarly emerge through iterative exchanges between contributors. Prompts can function as improvisational offers, while responses reflect forms of uptake, elaboration, adaptation, and reframing. Semantic shifts between conversational turns may therefore represent exploratory movement through conceptual space, whereas sustained elaboration and contextual responsiveness support the co-construction of ideas over time (Sawyer, 2021; Tang, 2025; Kumar et al., 2025). In this way, creativity within conversational collaboration extends beyond isolated outputs and instead emerges through the evolving interactional dynamics between contributors.

This interactional perspective on creativity is already reflected within evaluative frameworks used in embodied collaborative contexts, where creativity is examined through observable physical interaction between participants (Davis et al., 2017; Deshpande et al., 2023, 2024). Creative Sense-Making (CSM) and Observable Creative Sense-Making (OCSM) have primarily been applied to collaborative embodied activities such as pretend play, movement, and design interaction using expert human raters who code video-recorded interactional behavior. CSM conceptualizes creativity through shifts between unclamped states, where participants explore possibilities and generate new directions, and clamped states, where ideas are refined, stabilized, and evaluated (Davis et al., 2017). However, because these states require raters to infer underlying cognitive processes from observed interaction, the framework introduces interpretability and scalability challenges. OCSM addresses this limitation by shifting toward more directly observable characteristics of creative interaction through contribution-level dimensions such as participation, novelty, and appropriateness (Deshpande et al., 2023, 2024). Together, these frameworks provide an important foundation for examining creativity as an interactional and process-level phenomenon, motivating their extension into conversational human-LLM collaboration.

2.1.1. LLMs as co-creators

Contemporary LLM-based conversational systems increasingly reflect the characteristics of mixed-initiative interaction first introduced by Hearst et al., where humans and intelligent systems collaboratively shape interaction through reciprocal contribution and iterative refinement rather than rigid instruction alone (1999). Within conversational human-LLM collaboration, ideation develops through iterative exchanges in which contributors introduce, elaborate on, and respond to evolving ideas over time (Nomura et al., 2024; Geroimenko, 2026). Such interactions position LLMs not only as generation tools but also as co-creative collaborators participating in exploratory ideation and collaborative exchange (Geroimenko, 2026). As creativity emerges through reciprocal interaction between contributors, examining only final collaborative outcomes provides limited visibility into how individual contributions influence participation, exploratory idea development, and contextual grounding throughout interaction. In these contexts, both human and AI contributors reciprocally influence the direction, elaboration, and contextual grounding of collaborative ideation. Understanding creativity within human-LLM collaboration therefore requires examining individual contributions as they evolve throughout interaction rather than evaluating only final collaborative outputs. Such process-level evaluation becomes important not only for understanding how creative productivity emerges within collaboration, but also for examining whether both contributors participate meaningfully in ideation and whether LLM-based interaction supports or constrains human creative potential.

While recent studies have begun extending creativity evaluation beyond traditional divergent-thinking tasks into more interactive settings such as brainstorming and co-creation (Fukumura and Ito, 2025), these approaches continue to rely heavily on expert evaluation or final outcome assessment. Other work has sought to improve scalability through computational semantic measures of creativity (Hubert et al., 2024; Kumar et al., 2025); however, these methods largely evaluate completed outputs rather than the evolving process through which creative contributions emerge during collaboration. Together, these limitations motivate the need for scalable approaches capable of examining creativity at the level of individual contributions as they develop throughout interaction.

2.2. Extending Observable Creative Sense-Making to Text-Based Dialogue

Observable Creative Sense-Making (OCSM) is a process-level framework for evaluating interactional creativity through observable dimensions of creative contribution (Deshpande et al., 2023, 2024). In practice, expert human raters analyze collaborative interactions and score each participant's contributions across three dimensions: Participation, Novelty, and Appropriateness. Contributions are evaluated on a scale from 0 to 3 using predefined coding guidelines, producing detailed time-aligned trajectories that capture how creativity develops throughout interaction (Figure 1). By

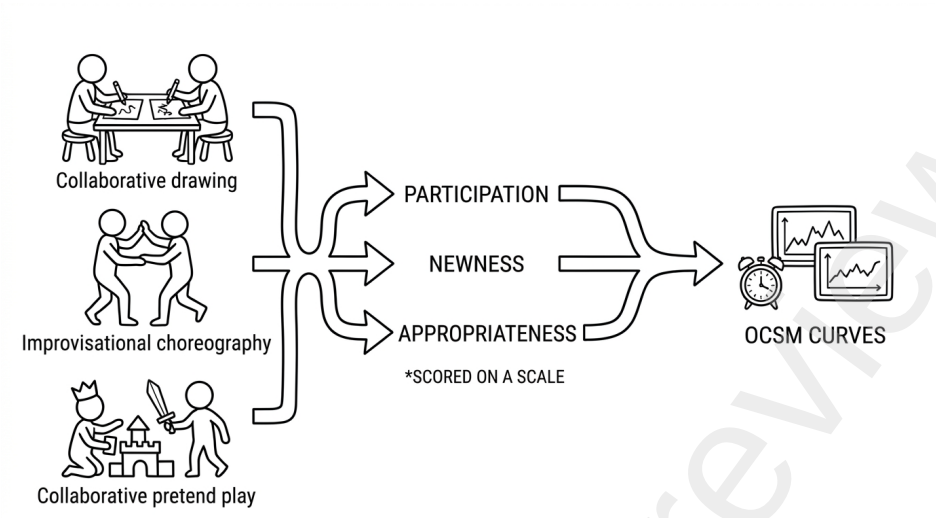

Figure 1: Overview of the Observable Creative Sense-Making (OCSM) framework for evaluating interactional creativity in collaborative tasks through participation, newness, and appropriateness over time.(Deshpande et al., 2023, 2024)

focusing on observable contribution dynamics, OCSM operationalizes interactional theories of collaborative creativity such as Sawyer's concept of "group genius," emergent product of interaction via the participation, responsiveness, elaboration, and the evolving transformation of ideas by the participants in the conversation by scoring each action by the participants on the following dimensions (Sawyer, 2017; Deshpande et al., 2023).

- **Participation** captures the level and type of engagement in the interaction, distinguishing between disengaged, independent, responsive, and co-creative contributions.
- **Novelty** captures the extent to which a contribution introduces new ideas, variations, or shifts in the interaction.
- **Appropriateness** captures how well a contribution aligns with the task and the evolving context of the interaction.

The coded observations are aggregated into temporal trajectories, enabling fine-grained analysis of how creative contributions evolve and interact over time. Participation captures mutual engagement and responsiveness between collaborators, while novelty and appropriateness reflect the introduction of new ideas alongside their contextual grounding within the evolving interaction.

To extend OCSM into text-based interaction, creative contributions must similarly be examined through sequential conversational exchanges rather than physical interactional behavior. Within dialogue, conversational turns function as observable units of participation, novelty, and contextual appropriateness that can be evaluated across interaction over time. Applying OCSM to conversational collaboration therefore enables the examination of how ideas are introduced, elaborated upon, responded to, and contextually maintained throughout Human–LLM interaction.Such an extension may also help address a persistent challenge in computational creativity evaluation: defining creativity in a theoretically grounded and operationally tractable manner (Kim et al., 2025). Rather than treating creativity as a singular construct, OCSM decomposes creative contribution into observable dimensions of participation, novelty, and appropriateness, providing a structured framework for evaluating creativity as it unfolds through interaction. By grounding creativity in theoretically derived multi-dimensional framework, OCSM offers a foundation for scalable process-level evaluation of creativity in dialogue.

2.3. Text Analysis Approaches to Operationalizing Interactional Creativity Dimensions

2.3.1. Understanding Computational Semantic and its Role in analyzing Novelty and Appropriateness

Computational semantics enables the analysis of text by representing words, sentences, or longer passages as vectors within a semantic space. The distance between these vectors indicates the degree of conceptual similarity, often quantified using cosine similarity measures (Green, 2016). Concepts with similar meanings are positioned close

together, whereas unrelated concepts are placed farther apart. For instance, "flying cars" and "sky highways" would be located near each other due to their association with futuristic transportation, while "sky highways" and "medical retraining" would be distant. This approach is grounded in distributional theories of meaning, which posit that language used in similar contexts conveys similar meanings. Early methods, such as Latent Semantic Analysis, analyzed word co-occurrence patterns to identify broad topic relationships but were limited in their ability to capture context (Heinen and Johnson, 2018; Alsaid et al., 2023). Subsequent embedding models, including Word2Vec and GloVe, advanced this field by learning word relationships from local textual context (Alsaid et al., 2023; Pezzotti et al., 2020).

More recently, sentence transformers have facilitated the generation of sentence-level embeddings that more effectively capture the meaning of extended ideas. For example, "we could build roads in the sky" and "imagine transportation happening above cities" are recognized as conceptually similar despite differing in wording (Mersha and Kalita, 2025). Researchers use these embedding-based methods to measure creativity in different fields. In tasks like the Alternate Uses Task (AUT), the semantic distance between answers shows how original and flexible the ideas are (Beaty et al., 2021). For longer tasks, such as writing stories, measures of semantic similarity and coherence check how well the content fits with what came before, showing its relevance and appropriateness (Fan et al., 2022; Singh et al., 2025). These studies suggest that semantic representations can capture both new ideas and how well they fit the context. Still, most research looks at single outputs instead of back-and-forth conversations.

To extend the OCSM as a scalable framework for evaluating creativity in text-based dialogue, two of the core dimensions are operationalized using this semantic technique: novelty and appropriateness. Novelty is measured as the semantic distance between conversational contributions, with greater distances indicating larger conceptual shifts and new ideas. For example, moving from discussing "improving roads" to proposing "airborne transit systems" is a novel shift. Appropriateness refers to how well each contribution aligns with the task or evolving dialogue context and can be assessed through semantic similarity to the prompt or previous conversational turns (Deshpande et al., 2023; Fan et al., 2022; Singh et al., 2025). For instance, suggesting "bike lanes" in a transport-planning discussion is highly appropriate, while introducing an unrelated topic is not. Together, these measures provide a scalable and mathematically rigorous framework for examining the dynamic unfolding of creativity in collaborative interactions.

2.3.2. Linguistic Measures in Text Analysis as an Indicator for Participation

Studying linguistic patterns to understand interactions began with early conversation analysis, where Sacks et al. showed that dialogue is organized through features like turn-taking, sequencing, and speaker contributions. This work proved that participation follows clear patterns, not random behavior (1974). Although their research was qualitative, it set the stage for computational methods that measure interaction at scale using language features. More recent studies have focused on measurable signs of participation such as word usage, punctuation, and engagement markers like reader-oriented pronouns ("we," "you") and directives ("let's try," "consider"). These markers show active involvement in group discussions (Busch, 2021; Seppänen et al., 2025; Pascual and Mur Dueñas, 2022). More verbal output is also linked to higher engagement in sharing information, serving as a quantitative sign of participation (Yoo and Kim, 2013). In line with OCSM's definition of participation as task engagement (Deshpande et al., 2023), and drawing on linguistic approaches that link verbal contribution and engagement markers to interactional involvement, we operationalize participation using two complementary measures. Word count captures contribution volume, and lexical engagement captures the extent of active involvement in the interaction. Together, these measures provide a clear and scalable way to quantify participation in both human–human and human–LLM interactions.

Moreover, applying these techniques to operationalize creativity within interaction may help address a second challenge identified in computational creativity research: the continued reliance on human evaluation for assessing creativity (Kim et al., 2025). By validating computational measures against expert ratings, this work explores a scalable approach to creativity assessment that remains grounded in human judgments of creative contribution.

2.4. Beyond Creative Performance: Human Centered Impact of Creative Collaboration

While this work focuses on creative contribution in interactional tasks, collaborative impact also depends on how participants experience the process. Prior research suggests that generative AI may improve creative efficiency while influencing users' autonomy, competence, and sense of meaning (Wei et al., 2025). Accordingly, a fuller human-centred account of collaboration should consider two complementary dimensions: subjective experience of the interaction and emotional change over the course of the task.

2.4.1. Human Experience in Interactional Creative Collaboration

Interactional creativity shows how people contribute, but it does not fully explain how humans experience collaboration. In creative work with LLMs, the subjective experience of the human teammate matters because it affects engagement, motivation, and whether someone trusts or questions their partner's input. Previous studies show that how people feel about their experience, confidence, and accountability can shape both their performance and participation in Human–AI teams (Yousefi et al., 2025). As AI systems become more like active collaborators instead of just tools, issues of autonomy, ownership, and responsibility become more important. When people feel more or less in control, it can change how they see their own role and contribution in a task (Biermann et al., 2022; Yousefi et al., 2025). Looking at subjective experience helps us understand not just what participants create, but also how they feel during the process.

Researchers often use self-report measures to study subjective experience in Human–AI creative work because qualities like confidence, ease, satisfaction, ownership, and accountability are personal and not directly visible. Self-reports work well in creative settings, where people's own sense of agency and meaning are key to understanding collaboration. In AI-assisted writing, earlier studies have used structured Likert-scale questions, based on workload and human factors research, to measure both how the process feels and how people view the results (Li et al., 2024). These questions cover enjoyment, how easy the task feels, the ability to express creative goals, the quality and uniqueness of the final work, pride, ownership, and willingness to take responsibility for issues like misinformation or harmful content (Li et al., 2024). This approach helps researchers systematically understand both the process and outcomes of collaboration, along with measures of interactional creativity.

2.5. Emotional Change and Creative Contribution

Emotional change is an important aspect of creative interaction because affective states can shape how people think, engage, and evaluate ideas during collaboration. Prior research shows that emotional valence and arousal influence key processes involved in creativity, including cognitive flexibility, associative thinking, and evaluative judgment (Ye et al., 2024). Emotional intensity and motivational state have also been linked to differences in creative engagement and output, although these effects may vary depending on individuals' baseline mood (Forgeard, 2011). In collaborative settings, changes in emotion may therefore influence not only how much people contribute, but also how they experience and sustain the creative process. To capture this dimension, affective responses are commonly assessed using dimensional models that distinguish between valence and arousal. The Self-Assessment Manikin (SAM) is a validated self-report measure widely used to assess these dimensions before and after task engagement (Bradley and Lang, 1994). Pre- and post-task SAM assessments provide a simple and established way to quantify emotional change associated with collaborative interaction (Ellard et al., 2011; Schock et al., 2012).

3. Method

3.1. Study Design

The present study employed a between-subjects design to compare Human–Human (HH) and Human–LLM (HL) brainstorming interactions using an identical creative task. The aim was to examine whether partner type influenced participants' creative contributions, subjective task experience, and emotional changes during collaboration. Participants were randomly placed in either the HH or HL group and took part in a structured brainstorming session with ten turns, taking turns with their partner. In both groups, they discussed the prompt: *"If all humans could fly starting tomorrow, how would this change cities, society, and daily life?"* This open-ended prompt was selected to encourage divergent thinking while maintaining a consistent and low-complexity task structure across participants, allowing us to isolate the effects of interaction type on creative processes. A set turn-taking was used to keep the interactions similar in both groups. Figure 2 shows an overview of the study procedure and analysis steps. Before the session, participants filled out surveys on demographics, personality (Big Five-Short), trust, cultural background, and their mood using the SAM Scale (Rammstedt et al., 2013; Frazier et al., 2013; Montag et al., 2023; Ellard et al., 2011). After the session, they completed surveys about their experience (overall process, confidence in the outcome, and accountability) (Li et al., 2024), and took the SAM Scale again to measure any changes in mood (Bradley and Lang, 1994). To measure creativity in the interactions, we used the OCSM framework, focusing on participation, novelty, and appropriateness. These were measured by analyzing the words and meanings in the conversations. To check the accuracy of these computer-based measures, expert raters also reviewed 22 conversations (11 HH and 11 HL) using rubrics based on the original OCSM guidelines to score the creative dimension as a collaborative dancing task (Deshpande et al., 2023).

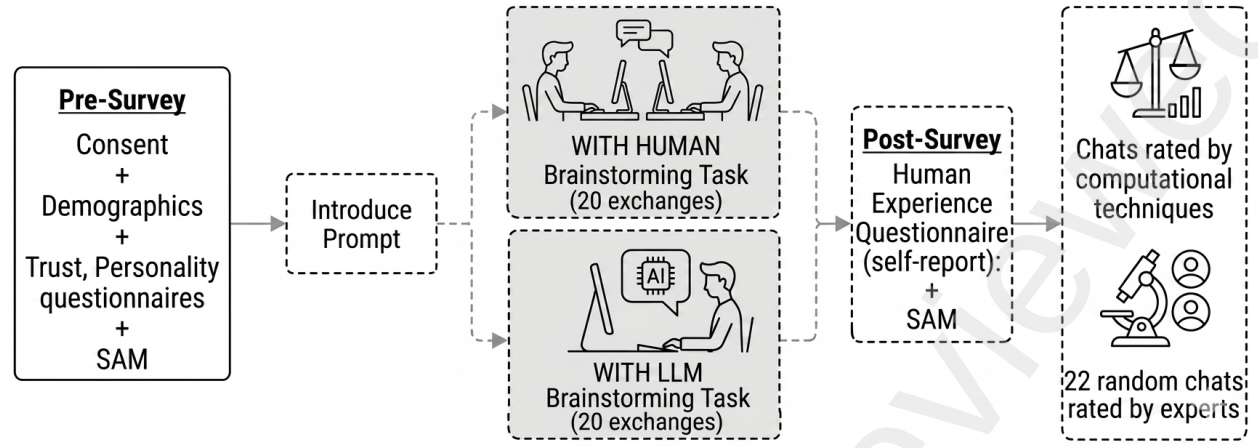

Figure 2: Experimental flow: pre-survey, brainstorming interaction (HH/HL), and post-task computational scoring with human validation.

3.2. Participants

Participants were recruited online using Prolific. To be eligible, they had to be at least 18 years old, live in the United States, and be fluent in English. In total, 105 people completed the study (ages 21 to 69, M = 41.3, SD = 11.87; 47.6% male, 51.4% female). Each participant received $12 as compensation. Participants were randomly placed into one of two groups: HH or HL. In the HH group, 40 participants were paired into 20 pairs. One person in each pair was randomly chosen as the main participant for analysis and started the interaction. In the HL group, 65 participants interacted one-on-one with an LLM partner. While the number of participants differed across conditions, the use of linear mixed-effects models enabled us to account for variability at the conversation level, providing robust estimates despite this imbalance. The study was approved by the Institutional Review Board at the University of Michigan–Dearborn (IRB #HUM00275518). All participants provided informed consent prior to participation through an online consent form.

3.3. Questionnaires

Participants completed brief pre- and post-task questionnaires to assess individual differences, baseline affect, and subjective task experience. Before the brainstorming task, participants completed demographic items, the short Big Five personality inventory (Rammstedt et al., 2013), general Propensity to Trust (PTT) and Propensity to Trust in Automation measures (PTT-A) (Montag et al., 2023), a short individualism-collectivism scale adapted from prior work (Wagner III, 1995), and a baseline Self-Assessment Manikin (SAM) measure. All questionnaire items used a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree), except the SAM, which used a 9-point pictorial scale.

Following the task, participants completed post-task measures of subjective experience and affect. Subjective experience items, adapted from prior work (Li et al., 2024), assessed process experience, outcome confidence, and accountability. Post-task affect was measured using the SAM (Bradley and Lang, 1994), and emotional change was calculated as the difference between pre- and post-task valence and arousal scores. This was done to account for any individual differences that could impact the collaboration performance or subjective experience.

3.4. Procedure and Experimental Treatments

Participants first completed a baseline survey assessing demographic characteristics and pre-task affect. They were then given standardized brainstorming instructions and completed a ten-turn (20 exchanges) structured dialogue in which turns alternated between partners, beginning with the focal participant and then the partner. In the HH condition, one participant within each dyad was randomly designated as the focal participant prior to the session and initiated the conversation. On the 9th turn of the focal participant, the dyad was prompted to begin concluding the discussion. Following the final exchange, participants completed post-task measures assessing affective change and subjective task experience.

In the **HL condition**, participants used a custom Streamlit interface accessed through a secure link. The brainstorming task appeared in a chat format similar to commercial AI chat systems. ChatGPT-3.5 was connected through an API and followed a standardized system prompt to ensure consistent interaction behavior across participants,

maintaining an inquisitive, constructive, and on-topic response style within the ten-turn structure (Ruangtanusak et al., 2025). In the **HH condition**, participants joined a scheduled Zoom session and did the same task using the chat function. A study moderator made sure participants took turns and gave the wrap-up prompt after the ninth exchange. The chat log feature was used to record the dialogue transcripts.

4. Results

Our main goal was to assess how interaction type and dialogue dynamics influence creativity. We used linear mixed-effects models (LMMs) to study interactional creativity, utterance-level participation, novelty, and appropriateness. This method handled repeated dialogue turns by adding random intercepts for conversation ID, capturing between-conversation differences. The fixed effects included interaction type (HH vs. HL), centered turn number, and their interaction, which let us compare groups and track changes over time. We analyzed utterances 1-19 to maintain consistent structures, since the last HL utterance was always generated by the LLM summarizing the entire chat.

To validate our computational creativity measures, we examined correlations between automated metrics and expert ratings, then used a decision tree to assess how well automated metrics predicted human-rated creativity. Exploratory analyses indicated that participant-level traits, including trust propensity, personality, and individualism–collectivism, showed weak and inconsistent associations with interactional creativity metrics across conditions, suggesting that observed differences were more strongly shaped by interaction type than by stable individual characteristics. We used ANOVA to analyze how interaction type influenced subjective experience and emotional change. Outcomes were measured once per participant, so we did not use repeated-measures models. Pearson's r explored links between individual differences and creativity outcomes. All analyses were conducted in Python using *pandas, statsmodels* libraries. Semantic representations were generated with the SentenceTransformers model *(all-MiniLM-L6-v2)*, which was used to derive embedding-based measures of novelty and appropriateness. All tests were two-tailed with $\alpha = .05$.

4.1. Participation

Participation captures the extent to which an individual engages with the task through their contributions to the dialogue, consistent with its definition in interactional creativity frameworks such as OCSM (Deshpande et al., 2023). We operationalized participation using two complementary measures: word score and lexical engagement. The *word score* was computed as the number of words per utterance (excluding stop words), normalized using quartile scoring (0–3) to reduce skew and enable comparability across conversations. This captures how much participants elaborate in each contribution or message, with higher word counts indicating greater engagement (Yoo and Kim, 2013). *Lexical engagement* was measured using linguistic features previously associated with engagement in collaborative discourse, to reflect the degree of active engagement expressed through linguistic cues.(Busch, 2021; Pascual and Mur Dueñas, 2022; Seppänen et al., 2025). These included punctuation (e.g., questions, exclamations), collaborative phrases (e.g., *"yes, and"*, *"yes, but"*), reader-oriented pronouns (e.g., *we, you*), and directive or modal language (e.g., *should, must, let's*). These features were extracted at the utterance level and aggregated into a composite score, which was similarly normalized into quartiles (0–3). Unlike word score, lexical engagement was computed on raw text to retain functional linguistic cues.

4.1.1. Participation Differences between HH and HL Participants

Participation varied significantly across interaction types. HL participants showed lower participation than HH participants across both measures. For word score, HL participants showed less elaboration per contribution (β = -0.574, $p = .001$), indicating shorter, less developed responses than HH interactions (Figure 3a). Similarly, for lexical engagement, HL interactions demonstrated lower use of interactional cues (β = -0.639, $p < .001$), reflecting reduced use of collaborative and directive language in shaping the dialogue (Figure 3b; Table 1). Participation also showed a significant decline over time for word score (β = -0.023, $p = .002$) (Figure 4a; Table 1), suggesting that contributions became shorter across turns as the interaction progressed. While a similar trend was observed for lexical engagement markers, this was not statistically significant (Figure 4b; Table 1). No significant interaction between group and turn was observed, indicating that the pattern of change over time was similar across both interaction types (Table 1).

4.1.2. Validation of Participation Measures

To validate the computational measures of participation, we compared them against human ratings provided by two expert annotators. Inter-rater reliability of participation scores was established on a pooled calibration set (36

Table 1
Linear mixed-effects model results for participation measures

	Word Score		Lexical Engagement Score	
	β	Std. Error	β	Std. Error
Intercept	1.231***	0.151	1.343***	0.142
HL vs. HH	-0.574**	0.173	-0.639***	0.162
Turn (centered)	-0.023**	0.007	-0.009	0.009
HL vs. HH x Turn	0.011	0.008	-0.007	0.011
Random intercept variance (Conversation ID)	0.423		0.344	
Residual variance	0.342		0.559	
Number of observations	848		848	
Number of conversations	85		85	
Log-likelihood	-867.052		-1049.285	

*** $p < .001$, ** $p < .01$, * $p < .05$

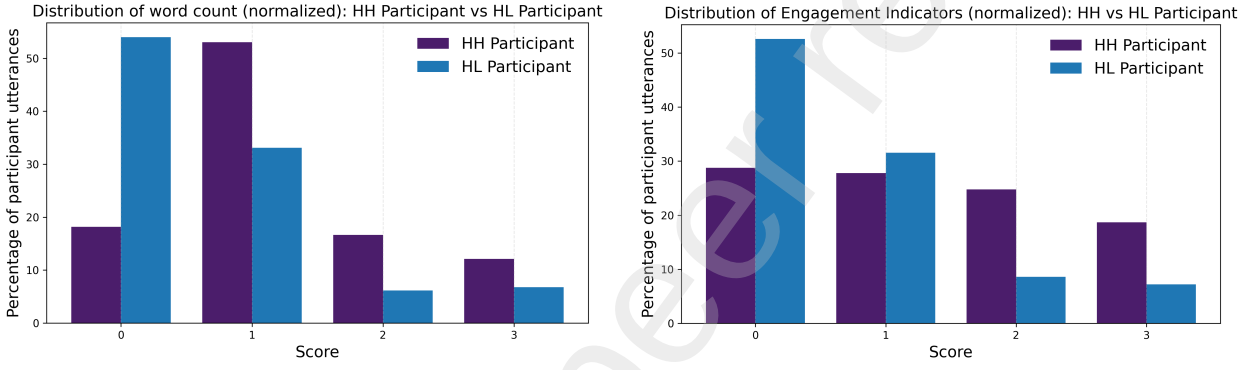

Figure 3: (a) Percentage distribution of word scores (0–3) across participant contributions in HH and HL interactions, showing shorter contributions in HL and more elaborated responses in HH. (b) Percentage distribution of lexical engagement scores (0–3) across participant contributions in HH and HL interactions, showing a higher concentration of low-engagement utterances in HL and higher engagement levels in HH.

utterances from two conversations), showing moderate agreement ($r = 0.60$, $\kappa = 0.53$). Following this, annotators independently rated participation across 11 HH and 11 HL conversations, evaluating all utterances within each dialogue in context. This differs from the group-level analyses above, which isolate individual participant contributions.Across all utterances, both word score and lexical engagement showed significant positive correlations with human-rated participation ($r = 0.53$ and $r = 0.49$, respectively, $p < .001$). A combined participation score further improved alignment ($r = 0.55$, $p < .001$), indicating that measures of elaboration and engagement capture complementary aspects of participation. Correlations were stronger in HL interactions ($r \approx 0.76$–0.80) than in HH interactions ($r \approx 0.17$–0.22), suggesting that participation in HL settings is more consistently reflected in linguistic patterns across both participants. Decision tree models further supported predictive validity. The combined participation score achieved the highest performance (accuracy = 71.4%, $\kappa = 0.50$), outperforming individual features.

4.2. Novelty

Novelty can be defined as the variance in ideas explored through collaboration, including the emergence of new ideas and changes in ideas as they evolve within a conversational context (Deshpande et al., 2023). In simple terms, it reflects how new or different a participant's contribution is relative to what has already been said. As discussed in Section 2.2, computational semantics enables the measurement of novelty through semantic distance, which has been widely used to capture originality in both divergent thinking tasks (e.g., AUT) and long-form text such as storytelling. Extending this approach to an interactional setting, we operationalize novelty as the semantic distance between a participant's current contribution (contribution which is being scored) and different reference points within the conversation. Each utterance at turn t, denoted as u_t, is represented as a sentence embedding vector.

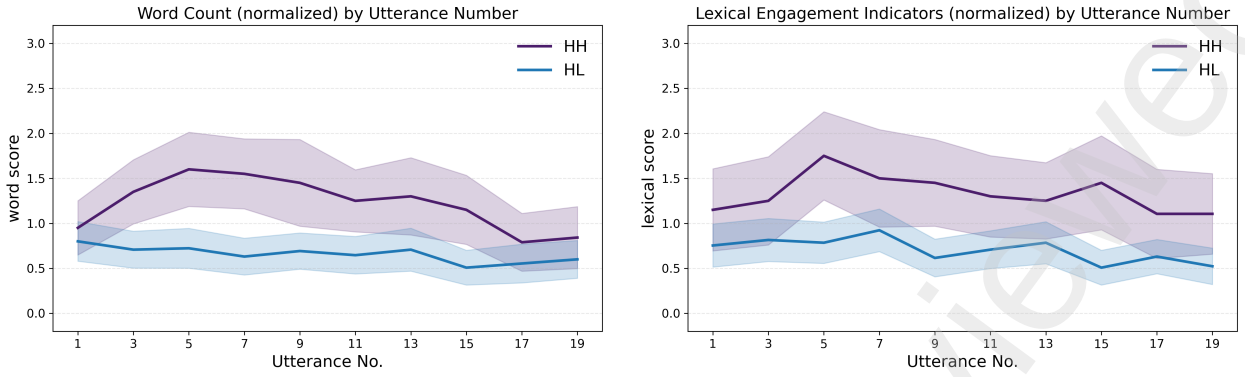

Figure 4: (a) Word score across dialogue turns for HH and HL participants. HH interactions show consistently higher scores, indicating greater elaboration per contribution, while both groups exhibit a gradual decline over time. Shaded regions indicate confidence intervals. (b) Lexical engagement across dialogue turns. HH interactions demonstrate higher use of interactional and engagement cues compared to HL, with relatively stable patterns across turns. Shaded regions indicate confidence intervals.

To quantify how much a contribution deviates from a reference utterance u_r, we compute cosine distance between their embeddings:

$$\text{Distance}(u_t, u_r) = 1 - \cos(u_t, u_r) \tag{1}$$

where u_t represents the current utterance and u_r represents a reference utterance from the conversation (1). Higher values indicate greater semantic divergence and were interpreted as higher novelty (Fan et al., 2022). Each utterance is converted into a sentence-level embedding using the *SentenceTransformer* model (*all-MiniLM-L6-v2*), which captures contextual meaning beyond word-level overlap. Novelty is then computed as the semantic distance between the current utterance (u_t) and a reference utterance (u_r), where u_r varies depending on the reference frame. Thus, for each contribution, we measure how far its meaning shifts relative to prior ideas in the conversation.

- **Overall Drift:** Measures how different a contribution is from the most similar prior idea in the conversation. Higher values indicate that the contribution introduces ideas that are novel relative to all previously expressed content.
- **Local Drift:** Measures how much a contribution diverges from the immediately preceding utterance. Higher values indicate a shift away from the partner's most recent idea (Figure 5a). This concept has been used to understand novelty in story writing tasks as Fan et al. computed distance between adjacent sentences and demonstrated this to be positively correlated with originality (2022).
- **Self Drift:** Measures how much a contribution differs from the participant's own previous contribution. Higher values indicate that the participant is introducing new ideas rather than repeating or refining earlier ones (Figure 5c).
- **Global Drift:** Measures how far a contribution moves away from the first idea at the start of the conversation (Figure 5b). Higher values indicate greater overall conceptual exploration. Previously, in an effort to gauge the quality of a writing task mean distance in the text has been shown positively correlate with originality (Fan et al., 2022).

4.2.1. Novelty Differences between HH and HL Participants

Novelty varied across interaction types and reference frames (Figure 6; Table 2). HL participants showed higher novelty at the local level ($\beta = 0.078$, $p = .004$), indicating that participants were more likely to move away from the immediately preceding idea (Figure 6a). HL participants also showed higher self drift ($\beta = 0.055$, $p = .013$), suggesting that participants introduced ideas that differed more from their own previous contributions (Figure 6b). Similarly,

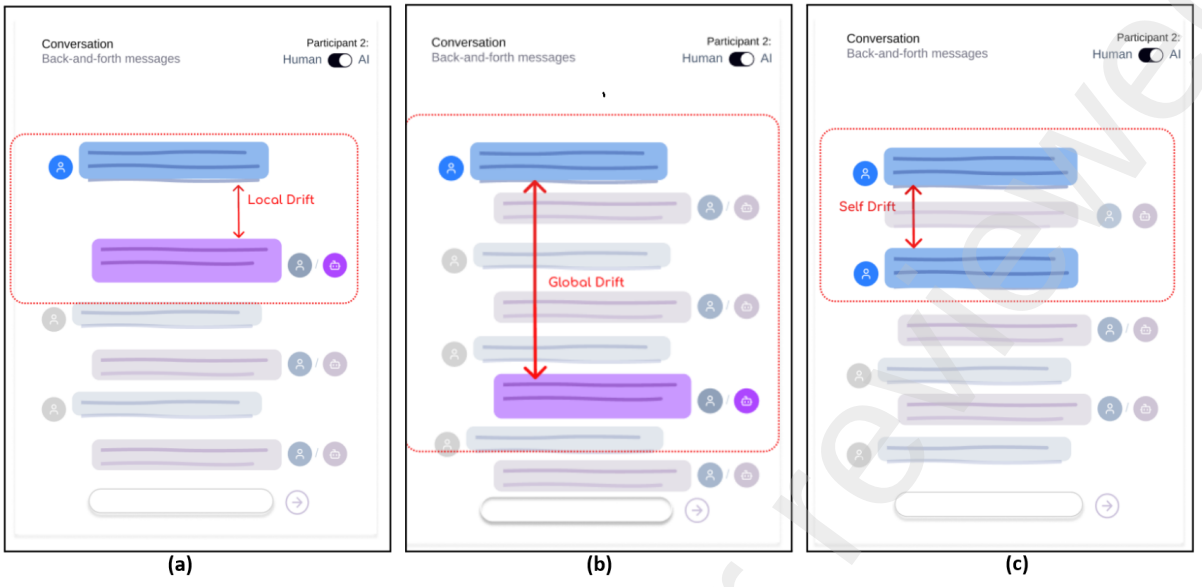

Figure 5: Conceptual representation of semantic drift reference frames. (a) **Local drift** captures how much a participant's current contribution diverges from their partner's immediately preceding idea. (b) **Global drift** captures how far a participant's contribution moves away from the initial idea introduced at the start of the conversation. (c) **Self drift** captures how much a participant's current idea diverges from their own previous contribution, reflecting intra-individual ideational expansion. Drift values were computed as cosine distance between sentence embedding representations.

Table 2
Linear mixed-effects model results for novelty (semantic drift measures)

	Local Drift		Self Drift		Global Drift		Overall Drift	
	β	Std. Error	β	Std. Error	β	Std. Error	β	Std. Error
Intercept	0.582***	0.024	0.616***	0.019	0.644***	0.024	0.489***	0.017
HL vs. HH	0.078**	0.027	0.055*	0.022	0.031	0.028	0.049*	0.020
Turn (centered)	0.005*	0.002	0.005*	0.002	0.008***	0.002	0.000	0.002
HL vs. HH x Turn	-0.001	0.002	0.000	0.002	-0.005*	0.002	-0.002	0.002
Random intercept variance	0.009		0.005		0.010		0.005	
Residual variance	0.018		0.020		0.015		0.014	
Observations	763		763		763		763	
Conversations	85		85		85		85	

*** $p < .001$, ** $p < .01$, * $p < .05$

overall drift was higher in HL participants ($\beta = 0.049$, $p = .015$), indicating that contributions were more distinct relative to prior ideas in the conversation (Figure 6c). However, no significant differences were observed between HH and HL participants for global drift (Figure 6d), indicating that both interaction types reached comparable levels of divergence from the initial idea. In contrast, global drift increased over time ($\beta = 0.008$, $p < .001$), indicating that conversations progressively moved further away from their starting point. This increase was weaker in HL interactions ($p = .021$), suggesting that HH interactions showed stronger cumulative exploration over time (Table 2; Figure 7). This suggests that while both interaction types reached comparable levels of divergence from the initial idea, the mechanisms through which this divergence emerged differed across conditions. Local drift also increased over time for participants in both the groups ($\beta = 0.005$, $p = .014$), indicating that participants increasingly shifted away from immediately preceding ideas as the interaction progressed (Table 2). No other effects were significant ($p > .05$).

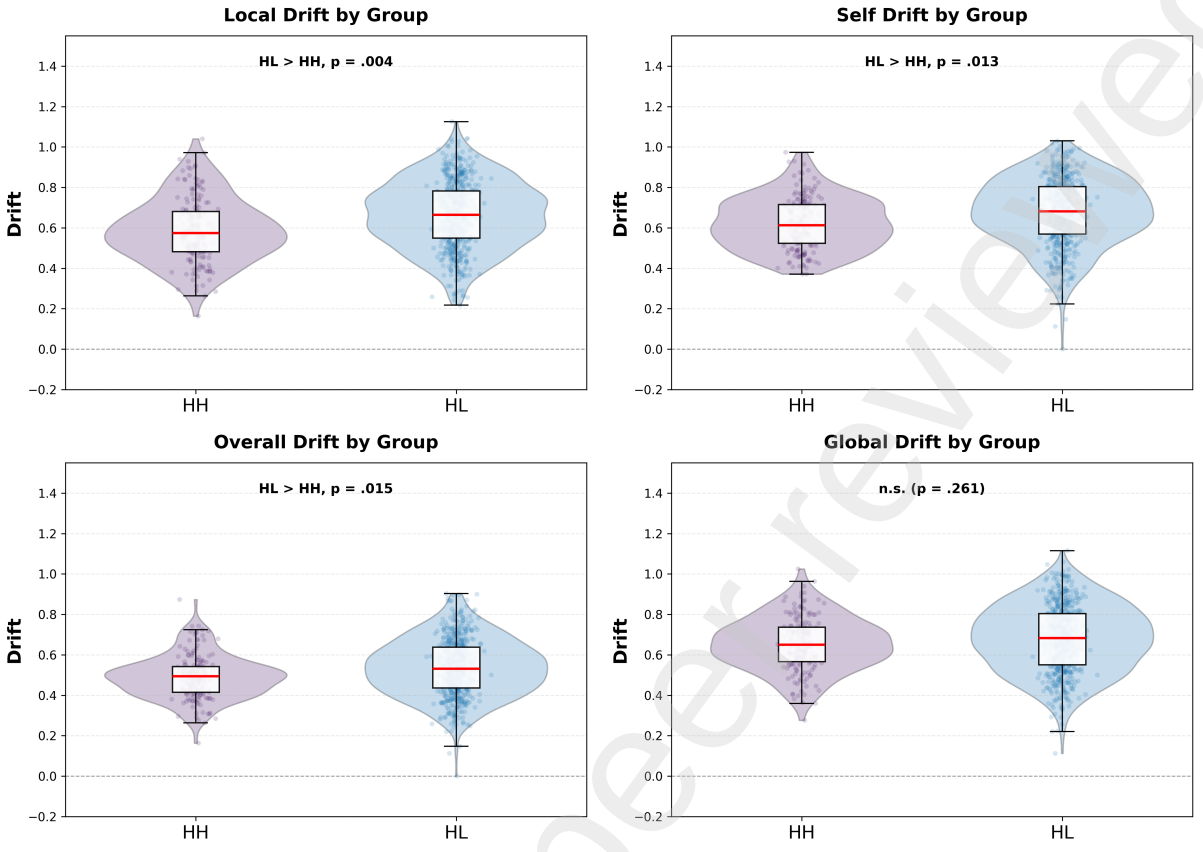

Figure 6: Distribution of novelty (semantic drift) across interaction types: Local, self, overall, and global drift. Local, self, and overall drift were consistently higher in HL participants showing more divergence of ideas were consistently higher in HL participants showing more divergence of ideas.

4.2.2. Validation of Novelty Measures

To assess construct validity, semantic drift measures were compared against human-rated newness scores. Inter-rater reliability for newness ratings was high ($r = 0.725$, weighted $\kappa = 0.711$), indicating consistency between expert raters. Correlation analyses revealed weak and inconsistent relationships between drift measures and human-rated newness. Across the full dataset, drift measures showed small negative or near-zero correlations with human ratings (e.g., self drift: $r = -0.131$, $p = .011$; global drift: $r = -0.103$, $p = .041$), with stronger effects observed primarily within HL interactions (e.g., self drift: $r = -0.219$, $p = .003$; global drift: $r = -0.165$, $p = .020$).

Decision tree models further showed limited predictive performance. A simple decision tree using drift features achieved low accuracy (mean accuracy = 0.35, weighted $\kappa = 0.01$), indicating minimal agreement with human-rated newness. More structured models (hierarchical and cross-validated trees) showed only marginal improvements (accuracy $\approx$ 0.40–0.46), but still exhibited weak agreement (weighted κ near zero), suggesting that drift measures do not reliably predict perceived novelty.

Across models, local drift emerged as the most informative feature, though its predictive contribution remained limited. These findings suggest that semantic drift does not directly align with human perceptions of novelty in conversational contexts. Rather than capturing perceived originality, drift measures reflect semantic divergence in embedding space. While prior work has validated semantic distance for novelty in structured tasks (e.g., AUT, long-form text), conversational settings involve additional dynamics that may not be fully captured by distance-based metrics. Accordingly, drift measures are better interpreted as indicators of semantic divergence within interaction, reflecting how ideas shift and evolve over time rather than perceived novelty alone.

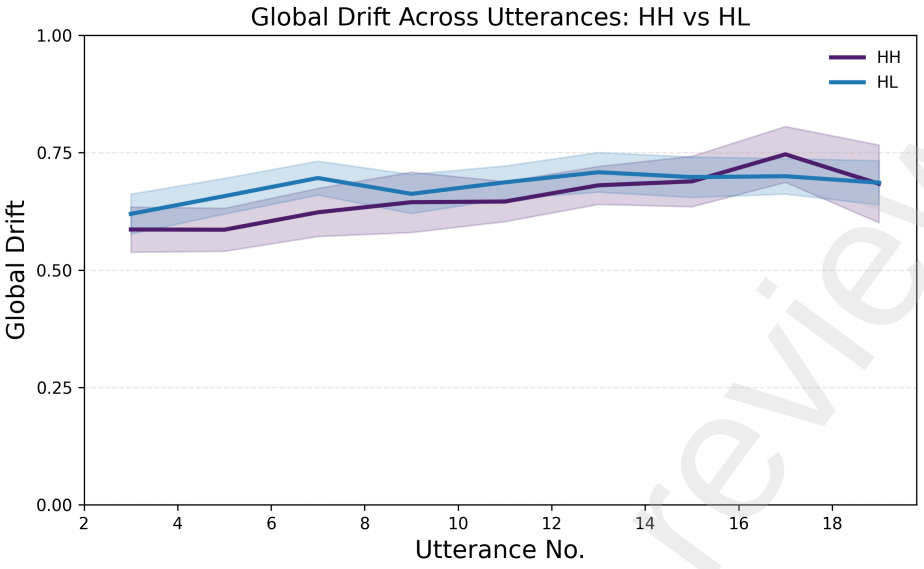

Figure 7: Global drift change over turns, with a stronger upward trend in HH interactions compared to HL ($p < .001$).

4.3. Appropriateness

Following the OCSM framework, **appropriateness** looks at the extent to which a participant's contribution remains relevant, coherent, and aligned with the task and evolving context of the interaction (Deshpande et al., 2023). In conversational settings, this reflects how well each utterance fits within both the original task and the ongoing dialogue. We operationalize appropriateness using semantic similarity between vector representations of utterances. Each utterance u_t at turn t is embedded into a vector representation $\mathbf{u}_t$ using a sentence embedding model. Semantic similarity between two utterances is computed as:

$$\cos(\mathbf{u}_t, \mathbf{u}_r) = \frac{\mathbf{u}_t \cdot \mathbf{u}_r}{\|\mathbf{u}_t\| \|\mathbf{u}_r\|} \tag{2}$$

where $\mathbf{u}_t$ is the embedding of the current utterance and $\mathbf{u}_r$ is the embedding of a reference. Higher values indicate greater semantic alignment (Equation 2). Building on prior work in narrative coherence, which evaluates how text relates to its broader context, we compute two forms of alignment (Singh et al., 2025; Fan et al., 2022). **Conversational alignment** is defined as the similarity between each utterance and the average embedding of the first three turns, capturing alignment with the early context of the interaction. **Task alignment** is defined as the similarity between each utterance and the original brainstorming prompt, capturing relevance to the task objective.

4.3.1. Appropriateness Differences between HH and HL Participants

Appropriateness varied across interaction types and decreased over time (3). HL interactions showed lower task alignment compared to HH interactions (β = -0.047, p = .035), indicating that contributions were less aligned with the original brainstorming prompt (Figure 8b; Table 3). Similarly, HL interactions exhibited lower contextual alignment (β = -0.047, p = .025), suggesting reduced alignment with the evolving conversational context (Figure 8a; Table 3).

Both task and contextual alignment decreased over time (task:β = -0.007, $p < .001$; context: β = -0.023, $p < .001$), indicating that contributions became progressively less aligned with both the task and prior context as the interaction progressed (Figure 9a,b). No significant interaction effects were observed ($p > .05$), indicating that the rate of decline over time was similar across HH and HL interactions (Table 3).

4.3.2. Validation of Appropriateness Measures

Validity for these metrics was tested against the scoring of the same conversation by expert raters who showed high inter rater reliability (r=0.927, κ=0.923) when scoring appropriateness, indicating a consistent benchmark for

Table 3
Linear mixed-effects model results for appropriateness measures

	Task Alignment		Contextual Alignment	
	β	Std. Error	β	Std. Error
Intercept	0.388***	0.020	0.501***	0.018
HL vs. HH	-0.047*	0.022	-0.047*	0.021
Turn (centered)	-0.007***	0.002	-0.023***	0.002
HL vs. HH x Turn	0.000	0.002	0.002	0.002
Random intercept variance	0.006		0.004	
Residual variance	0.017		0.023	
Observations	848		848	
Conversations	85		85	
Log-likelihood	451.724		334.854	

*** $p < .001$, ** $p < .01$, * $p < .05$

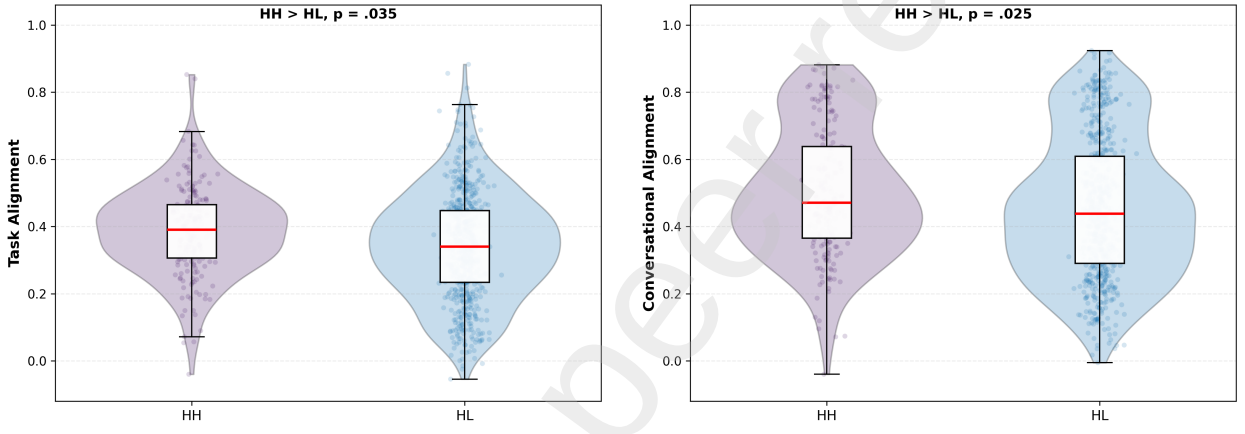

Figure 8: Distribution of appropriateness across interaction types. (a) Conversational alignment and (b) task alignment for HH and HL interactions. HL interactions show consistently lower alignment than HH, indicating reduced coherence with both the evolving conversation and the original task.

validation. Construct validity was supported by significant positive correlations between computed alignment measures and human-rated appropriateness. Task alignment showed the strongest association overall ($r = 0.417, p < .001$), followed by conversational alignment ($r = 0.320, p < .001$). Both measures were positively correlated with human ratings across HH and HL subsets.

Predictive validity was further assessed using decision tree models. A simple model based on task alignment achieved moderate performance (accuracy = 63.83%, weighted $\kappa = 0.233$). Hierarchical models showed improved performance, with the combined alignment score emerging as the best predictor (accuracy = 70.83%, weighted $\kappa = 0.316$), followed by task alignment alone (accuracy = 69.83%, weighted $\kappa = 0.295$). These results suggest that semantic alignment measures provide a meaningful approximation of human-perceived appropriateness, with task relevance playing the strongest role.

4.4. Subjective Experience

Human experience was assessed to complement interactional creativity by capturing how participants perceived the collaboration process and its outcomes. Prior work highlights its role in shaping engagement, outcome evaluation, and AI adoption (Yousefi et al., 2025). We measured subjective experience using structured self-report items adapted from (Li et al., 2024), rated on a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). Three dimensions were measured:

1. **Overall Experience** (4 items): capturing satisfaction with the brainstorming process, enjoyment, perceived ease, and ability to express creative goals;

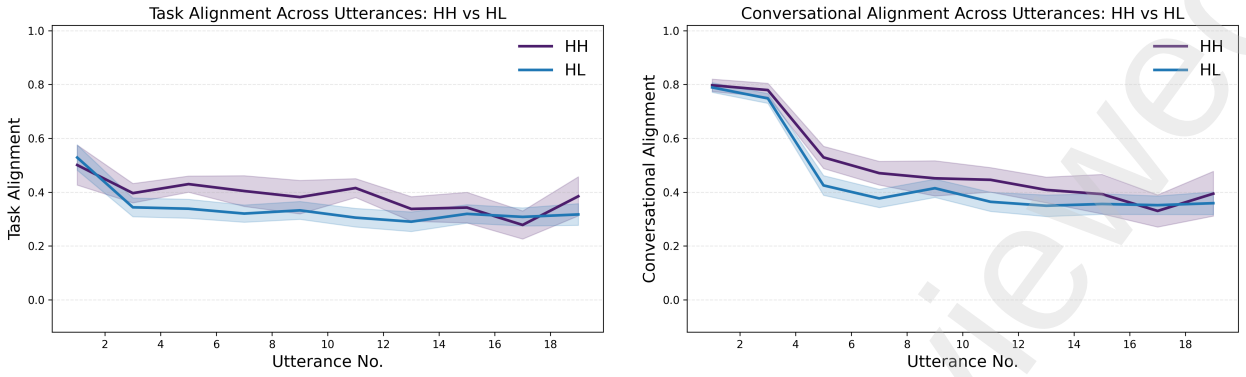

Figure 9: Appropriateness change across interaction turns.(a) Task alignment and (b) conversational alignment across utterances for HH and HL interactions. Both measures decrease over time, indicating reduced alignment with the task and conversational context as interaction progresses, with consistently lower alignment observed in HL compared to HH.

Table 4
Group differences in human experience measures

Measure	HH participant	HH partner	HL participant	ANOVA F	ANOVA p	Kruskal p
Overall Experience	4.51 (0.66)	4.64 (0.60)	4.27 (0.86)	2.05	.134	.104
Perceived Outcome Confidence	4.51 (0.56)	4.63 (0.55)	4.12 (0.86)	4.42*	.014	.005
Accountability	4.36 (0.89)	4.53 (0.71)	4.19 (0.97)	1.14	.324	.240

Values are reported as mean (SD).

Table 5
Post hoc comparisons for perceived outcome confidence

Comparison	Mean Difference	Adjusted p	Significant
HH participant vs. HH partner	0.113	.887	No
HH participant vs. HL participant	-0.393	.113	No
HH partner vs. HL participant	-0.506	.029	Yes

2. **Perceived Outcome Confidence** (4 items): capturing satisfaction with the final outcome, perceived quality, ownership, pride, and perceived uniqueness;
3. **Accountability** (4 items): capturing willingness to assume responsibility for potential shortcomings in the final output (e.g., misinformation, plagiarism, privacy invasion, bias).

Items were adapted to reflect the brainstorming task context while preserving the original conceptual structure. Responses were averaged within each dimension to compute composite scores used in analysis.

4.4.1. Differences in Human Experience

Human experience differed only for perceived outcome confidence (Table 4). The omnibus effect was significant for perceived outcome confidence, $F = 4.42, p = .014$, and was also supported by the non-parametric Kurksal test ($H = 10.61, p = .005$), to account for any normality violations. Post hoc comparisons showed that HL participants reported lower perceived outcome confidence than HH partners ($p = .029$), while differences between HL participants and HH participants were not significant, and no difference was observed between the two HH roles (Table 5). Individual differences showed minimal associations with subjective experience measures, suggesting that perceived workload, confidence, and overall experience were primarily influenced by the interaction context rather than participant traits.

4.5. Emotional Change

Emotional change was defined as shifts in affective state, operationalized as the difference in valence (pleasantness) and arousal (activation) before and after the task, which together capture the core dimensions of emotional experience

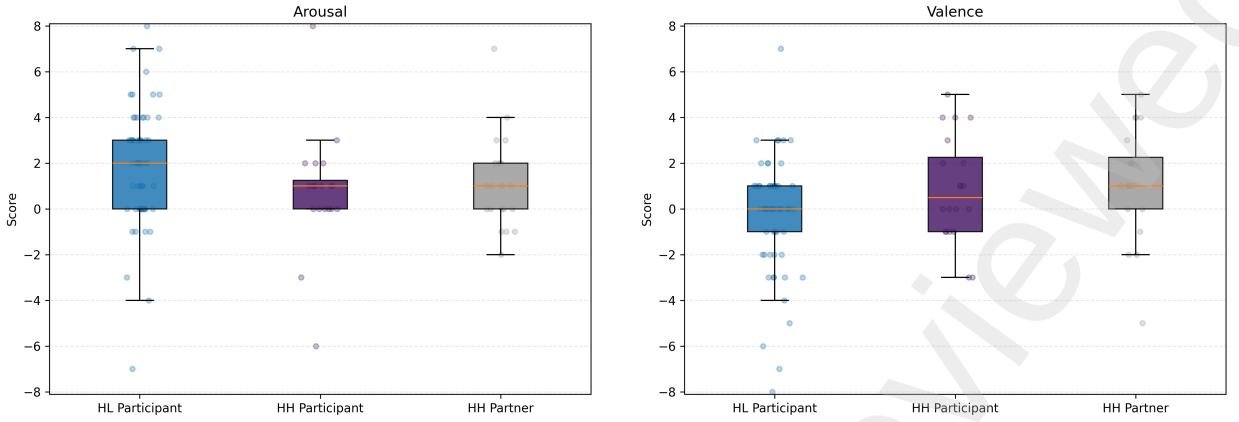

Figure 10: Emotional change across partner groups (HH participant, HH partner, HL participant). Boxplots display post–pre difference scores for (a) ΔArousal and (b) ΔValence (Self-Assessment Manikin scale). Positive values indicate increases following interaction.

(Ellard et al., 2011; Bradley and Lang, 1994). Participants reported their affective state before and after the task on a 9-point non-verbal pictoral SAM scale, and emotional change was computed as the difference (post–pre) in valence and arousal, reflecting shifts in emotional experience over the course of the interaction.

4.5.1. Differences in Emotional Change

Emotional change differed across interaction types for valence, with a significant main effect of group ($F(2, 82) = 3.58, p = .032$). Comparisons between HH and HL participants indicated a trend toward higher valence change in HH participants ($p = .163, n.s.$) (Figure 10a,b). This pattern was more pronounced when considering interaction partners, with HH partners showing higher valence change than HL participants ($p = .057$, marginal), suggesting relatively less positive affect in human–LLM interactions. No other group differences were observed. In contrast, no significant effect of group was found for arousal ($F(2, 82) = 1.95, p = .148$). Exploratory analyses indicated that emotional change was not consistently predicted by interactional creativity metrics ($p > .05$). Changes in affect were also not meaningfully explained by baseline individual differences, indicating that emotional shifts were more likely driven by the collaborative experience itself.

5. Discussion

This study examined how collaborative creative contributions differ when people work with a large language model rather than another person. The researchers introduced a process-level framework that evaluates creativity by analyzing each turn in a conversation. Creativity of each contribution was defined by three observable characteristics: participation, novelty, and appropriateness. This approach enabled tracking how contributions evolve and change throughout the conversation, rather than just looking at the final results.

We used computational analysis of conversation data to measure these aspects at each turn. The measures for participation and appropriateness matched expert human ratings, which supports their validity. Novelty was not directly validated, but was measured using semantic divergence, a method from earlier studies to capture differences in ideas (Beaty et al., 2021; Fukumura and Ito, 2025; Fan et al., 2022; Mersha and Kalita, 2025; Singh et al., 2025; Green, 2016; Heinen and Johnson, 2018). This shows a common challenge in creativity research: creativity is subjective and often debated, with definitions of originality and value varying across fields and viewpoints (Kaufman and Sternberg, 2010). By combining validated dimensions with established computational methods, this framework stays aligned with human judgments where possible and remains scalable. It also allows for structured comparisons between partner types and interaction contexts, helping to evaluate collaborative performance, find design gaps, and benchmark human–AI systems.

5.1. Interactional Creativity Shifts Systematically Across Partner Types

5.1.1. LLM Collaboration Reduces Participation and Engagement

Participation is a key part of interactional creativity. It shows how people take part in the co-creative process and how ideas are shared, accepted, and developed together (Deshpande et al., 2023; Davis et al., 2017; Sawyer, 2021; Smuts, 1992). While creativity is often defined by novelty and usefulness, which match originality and appropriateness in this study (Kaufman and Sternberg, 2010), participation adds to these ideas by focusing on how much and in what ways people join in the collaborative process.

Participants who worked with LLMs showed lower participation than those who worked with another person. This was seen in shorter messages and less varied language. These measures show not just how much people contribute, but also how involved they are in the interaction and how actively they engage with the task and their partner. Higher participation in human-human interactions suggests more sustained elaboration, where individuals actively build on ideas, respond to their partner, and contribute to the development of the discussion. Although participation declined across turns in both conditions, this pattern likely reflects the natural progression of structured brainstorming tasks, in which initial exploration gives way to consolidation. However, the consistently lower participation observed in the human-LLM condition indicates reduced engagement in the co-creative process.

From a participatory perspective, these findings speak directly to the questions raised by Hanchett Hanson (2015), who emphasize that creativity is not only about generating new ideas but also about how individuals contribute, take up roles, and shape interactions through their engagement. In human-human collaboration, these roles appear to be more dynamically shared, with participants alternating between generating, elaborating, and refining ideas. In contrast, human-LLM interaction may involve a more asymmetric distribution of roles. Prior work suggests that LLMs often take on roles in elaboration, suggestion, and idea evaluation (Shaer et al., 2024), which may reduce the extent to which human participants engage in these processes themselves. As a result, participants interacting with an LLM may contribute less actively to the development of ideas, leading to lower participation despite the presence of diverse ideas within the interaction.

5.1.2. LLM Collaboration Increases Semantic Divergence but Limits Conceptual Progression

Applying this framework revealed systematic differences in how human creative contributions developed when collaborating with another human versus a large language model. Participants in the human–LLM condition consistently demonstrated higher semantic divergence than those brainstorming with another human. This pattern was observed across local, self, and overall measures of divergence, indicating that participants introduced ideas that differed not only from their immediate prior turn (local), but also from their own earlier contributions (self) and the broader conversational context (overall). Together, these patterns suggest that interaction with an LLM encouraged broader exploration of the semantic space, prompting participants to generate ideas that were increasingly varied relative to both their own and their partner's prior contributions. These findings align with prior work showing increased novelty in human–LLM collaboration at the outcome level (Nomura et al., 2024; Hubert et al., 2024). One possible explanation is that LLMs introduce unexpected ideas or associations that prompt users to explore new directions, as demonstrated in prior work where LLM-generated content inspired users to consider ideas they may not have otherwise generated (Lin et al., 2025). In the present study, this influence appears to extend to the interaction process itself, where participants not only respond to such inputs but also generate increasingly diverse contributions relative to both their own and the ongoing conversational context.

Within the human-centered design domain, prior work has shown that while LLM-assisted collaboration does not necessarily produce more original or higher-quality outcomes, it enables faster ideation and reduces the time required to explore design alternatives (Zhou et al., 2024). This challenges the assumption that increased novelty in human–LLM collaboration directly translates to improved outcomes. One possible explanation is that observed gains in novelty may reflect semantic divergence rather than substantive originality, leading to outputs that appear diverse but are qualitatively similar; this may also explain the lack of alignment between human raters' scores and our computed novelty scores. Together, these findings suggest that human–LLM interaction is particularly well-suited for early divergent phases of the creative process, where rapid exploration of ideas is valuable.

Global divergence did not differ significantly between interaction types at the aggregate level; however, temporal analyses revealed important differences in how divergence evolved over time. Human–human interactions exhibited a stronger increase in global divergence across turns, indicating greater cumulative movement away from the initial idea. This suggests that while both interaction types ultimately reached comparable levels of divergence, they did so through different mechanisms. Human–LLM interactions promoted divergence at the level of individual contributions, whereas

human–human interactions supported more sustained, cumulative development of ideas across the interaction. Given that global divergence was computed as the distance of each contribution from the initial message of the conversation, this pattern suggests that, while ideas introduced during human–LLM interaction varied across successive turns, these contributions remained relatively anchored to the original framing. In other words, although participants frequently generated new ideas, these contributions were less likely to collectively shift the discussion's overall direction. In contrast, those working with a human peer appeared to engage in a more cumulative form of idea development, in which contributions built on one another to reshape the discussion over time. Rather than introducing isolated variations, participants extended, refined, and reinterpreted prior ideas, enabling the interaction to evolve toward new conceptual directions. This distinction highlights a key difference in how novelty manifests across conditions.

Previous work has demonstrated increased novelty in human–LLM collaboration using outcome-based measures, including the number of ideas generated (Nomura et al., 2024) and performance on standardized creativity tasks (Kumar et al., 2025; Hubert et al., 2024; Beaty et al., 2021). These findings align with the higher local, self, and overall divergence observed in the present study, reflecting increased variation in ideas at the level of individual contributions. However, such measures primarily capture the expansion of ideas rather than the development and integration of ideas over time. The present findings extend this work by showing that, despite generating a wider range of ideas, contributions in human–LLM interactions were less likely to collectively shift the direction of the conversation. In contrast, human–human interactions supported a more cumulative evolution of ideas across turns. This interpretation aligns with prior findings suggesting that, although human–LLM collaboration can accelerate ideation, the resulting outputs may remain qualitatively similar to those produced through human collaboration (Zhou et al., 2024), indicating that the speed of idea generation does not necessarily translate into deeper conceptual evolution.

5.1.3. Increased Divergence is Accompanied by Reduced Appropriateness

The divergence–appropriateness tradeoff observed in the data suggests a predictable tension between ideational expansion and coherence. As divergence increases, particularly in human-LLM interactions, contributions become less grounded within the shared conversational context. A general decline in appropriateness across turns in both conditions further reflects the difficulty of maintaining alignment as ideas expand. However, the consistently lower appropriateness observed in human-LLM interactions indicates a more pronounced manifestation of this trade-off. This finding aligns with prior work on discursive coherence in human–AI interaction, which shows that AI systems can produce coherence along dictional, intentional, emotional, and rational dimensions, yet lack the structural depth and generative integration observed in human-human dialogue (Tang, 2025). As a result, while LLM contributions may remain locally coherent, they may not support the cumulative development of ideas, prompting participants to introduce new directions more frequently and contributing to the higher levels of divergence observed in the human–LLM condition. This pattern may also reflect differences in role distribution within the interaction. While human collaborators tend to share responsibilities for generating, elaborating, and integrating ideas, LLMs more often take on roles related to suggestion and surface-level elaboration, leading human participants to assume a more active role in introducing new directions rather than collectively developing existing ideas.

In summary, participants who interacted with another human demonstrated a more balanced distribution of creative contributions compared to those engaging with a large language model. These participants combined the generation of new ideas with sustained engagement and strong contextual alignment. This pattern reflects a more integrated co-creative process, in which ideas are not only introduced but also developed, negotiated, and grounded within the interaction. In contrast, human–LLM interactions exhibit a less balanced distribution of creative roles. Participants engaging with a large language model generated a greater variety of ideas, both individually and collectively, but participated less frequently and produced contributions that were less contextually appropriate than those of participants interacting with a human peer. These findings indicate that, while collaboration with an LLM supports the generation of diverse ideas, it does not facilitate the same level of integration or collaborative development observed in human–human interaction. Rather, the interaction emphasizes variation over the co-construction of ideas.

Notably, human participants who interacted with LLMs continued to make creative contributions. The results suggest that LLM interactions support specific aspects of creativity in human participants, particularly originality and rapid idea generation, but do not equally encourage other dimensions of creativity. This emphasis may be advantageous for brainstorming or early-stage ideation, but appears less supportive of processes such as refinement, collaborative development, and shared ownership, which are more evident in human–human interactions. Importantly, these patterns were not meaningfully explained by individual differences. Exploratory analyses of trust, personality, and individualism-collectivism revealed weak and inconsistent associations, indicating that the structure of the interaction

itself played a more central role in shaping creative contributions than participant-level traits. Overall, human–LLM collaboration is best understood as a complementary mode of interaction that redistributes creative effort across different dimensions, rather than uniformly enhancing creativity. This observation highlights a gap in current systems and suggests the need for LLMs that better support holistic creative development.

5.2. Human–LLM Collaboration Alters Perceived Outcomes and Emotional Experience

Human-centered outcomes showed clear differences between types of partners. Although overall experience and accountability were similar for both groups, participants who worked with a peer felt more confident about the results than those who worked with a large language model. Importantly, none of the experience measures were linked to creativity metrics, so more creative or active participation did not lead to better perceived quality. Instead, people's perception of outcomes or confidence in the quality seemed to depend more on who they interacted with than on how much each person contributed. This result differs from earlier studies that found higher perceived outcomes when using LLMs, which may be attributed to differences in task type, interaction balance, or offered incentives (Li et al., 2024). Earlier studies used tasks with clear rewards, whereas the present study used a more open-ended brainstorming task with lower stakes, which may result in differences in the expected quality of the outcome.

Accountability was similar in all conditions, which suggests that lower perceived outcomes in human-LLM interactions were not caused by feeling less responsible. Instead, these differences may come from how much people feel they share ownership and control during the interaction. Working with another person can create a stronger sense of shared authorship and investment, while working with a large language model may make joint contribution feel less real. This idea matches earlier research on autonomy and agency in human-AI teamwork (Bangerl et al., 2025; Biermann et al., 2022).

Emotional responses also set the interaction types apart. While arousal levels were about the same, people who worked with another person felt more positive, while those who worked with a large language model felt slightly less positive on average. Like the experience measures, changes in emotion were not linked to creativity, which suggests that feelings were not directly shaped by how creative or active people were. Instead, emotions seem to be shaped by the social side of the interaction. Working with another person can offer feedback, encouragement, and social support, which boost positive feelings. In contrast, working with a large language model, even though it helps generate ideas, may not provide the same social cues.

These results show that creative contribution, personal evaluation, and emotional experience are partly separate aspects of working together. Human-LLM interaction helps people come up with more ideas, but it does not always lead to better perceived results or more positive feelings. Earlier research suggests that feeling positive is important for keeping creativity and engagement high over time, so the lower positive feelings seen in human-LLM interactions could affect long-term creative work. Overall, these findings suggest that when judging human-AI systems, it is important to look at both what is produced and how people feel during the process.

6. Implications

This study has important implications for both theory and practice in designing and evaluating human–AI interaction. On the theoretical side, it introduces a process-focused way to study creativity in collaborative settings, making it possible to see individual contributions during an interaction. This approach moves beyond just looking at outcomes and instead examines how creativity develops as people work together, offering a clearer picture of the creative process. On the practical side, the findings can help guide the design of conversational systems that support human creativity. The results show that working with LLMs encourages people to generate a wider range of ideas, but it may also lead to less ongoing participation and less integration of those ideas. This means that conversational agents should not only generate new responses but also help keep users engaged, maintain continuity, and support building ideas together. For instance, systems could prompt users to build on earlier ideas, keep track of the conversation, or revisit previous contributions to help ideas grow over time. The framework introduced here also makes it possible to evaluate creativity in real time, providing a practical way to see how conversational systems affect user behavior. This kind of evaluation can help compare different models and designs, both for their individual performance and for how well they work as partners in collaboration. This is especially useful for HCI applications like creative tools, educational platforms, and collaborative systems, where it is important to understand how technology shapes human input. These findings also affect how people use AI in creative settings. Different ways of interacting highlight different aspects of creativity: working with LLMs encourages a variety of ideas, while working with other people supports combining

ideas and staying engaged. Knowing about these differences can help users choose the right tools and strategies for their creative goals, and it shows why it is important to design systems that support, rather than replace, human creativity.

7. Conclusion

This study examined how collaboration with a large language model influences human creative contributions using a process-level perspective. By extending the OCSM framework to text-based interaction and operationalizing creativity at the level of individual conversational turns, this work provides a scalable approach to understanding how creativity develops during collaboration. Findings show that human–LLM interaction redistributes creative effort across dimensions, increasing semantic divergence while reducing participation, appropriateness, and cumulative development of ideas. In contrast, human–human interaction supports more sustained engagement and integrative idea development over time. These results suggest that the impact of AI on creativity is not uniform, but depends on how interaction shapes different aspects of the creative process. From an HCI perspective, this highlights the importance of designing conversational systems that not only support idea generation but also facilitate engagement, context maintenance, and collaborative development. As LLMs become increasingly embedded in interactive systems, understanding how they influence human behavior is essential for designing experiences that enhance, rather than fragment, creative processes. By enabling scalable, process-level evaluation of creativity in interaction, this work provides a foundation for designing and benchmarking conversational systems that support balanced human–AI collaboration. Future research should extend this framework across more complex tasks, diverse interaction settings, and evolving model capabilities to further inform the design of effective and human-centered AI systems.

Declaration of Generative AI Use

The authors acknowledge the use of ChatGPT (OpenAI, GPT-4) and Grammarly AI (Grammarly Inc.) for language editing and refinement during the preparation of this manuscript. All outputs generated using these tools were carefully reviewed and edited by the authors, who take full responsibility for the content of this manuscript.

Disclosure Statement

The authors report no potential conflicts of interest. The authors alone are responsible for the content and writing of this article.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author, A. Alsaid, upon request.

References

Alsaid, A., Li, M., Chiou, E.K., Lee, J.D., 2023. Measuring trust: a text analysis approach to compare, contrast, and select trust questionnaires. Frontiers in Psychology 14. URL: http://dx.doi.org/10.3389/fpsyg.2023.1192020, doi:10.3389/fpsyg.2023.1192020.

Baldeo, S., 2026. Generative artificial intelligence reliance and executive function attenuation: Behavioral evidence of cognitive offload in high-use adults. Technology, Mind, and Behavior .

Bangerl, M.M., Disch, L., David, T., Pammer-Schindler, V., 2025. Creaitive collaboration? users' misjudgment of ai-creativity affects their collaborative performance, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM. p. 1–17. URL: http://dx.doi.org/10.1145/3706598.3713886, doi:10.1145/3706598.3713886.

Barron, F., Harrington, D.M., 1981. Creativity, intelligence, and personality. Annual review of psychology .

Beaty, R.E., Zeitlen, D.C., Baker, B.S., Kenett, Y.N., 2021. Forward flow and creative thought: Assessing associative cognition and its role in divergent thinking. Thinking Skills and Creativity 41, 100859. URL: http://dx.doi.org/10.1016/j.tsc.2021.100859, doi:10.1016/j.tsc.2021.100859.

Biermann, O.C., Ma, N.F., Yoon, D., 2022. From tool to companion: Storywriters want ai writers to respect their personal values and writing strategies, in: Designing Interactive Systems Conference, ACM. p. 1209–1227. URL: http://dx.doi.org/10.1145/3532106.3533506, doi:10.1145/3532106.3533506.

Bradley, M.M., Lang, P.J., 1994. Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25, 49–59. URL: http://dx.doi.org/10.1016/0005-7916(94)90063-9, doi:10.1016/0005-7916(94)90063-9.

Bryan-Kinns, N., Healey, P.G.T., Leach, J., 2007. Exploring mutual engagement in creative collaborations, in: Proceedings of the 6th ACM SIGCHI conference on Creativity & cognition - C&C '07, ACM Press. p. 223. URL: http://dx.doi.org/10.1145/1254960.1254991, doi:10.1145/1254960.1254991.

Busch, F., 2021. The interactional principle in digital punctuation. Discourse, Context & Media 40, 100481. URL: http://dx.doi.org/10.1016/j.dcm.2021.100481, doi:10.1016/j.dcm.2021.100481.

Davis, N., Hsiao, C.P., Singh, K.Y., Lin, B., Magerko, B., 2017. Creative sense-making: Quantifying interaction dynamics in co-creation, in: Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition, ACM. p. 356–366. URL: http://dx.doi.org/10.1145/3059454.3059478, doi:10.1145/3059454.3059478.

Deshpande, M., Park, J., Pait, S., Magerko, B., 2024. Perceptions of interaction dynamics in co-creative ai: A comparative study of interaction modalities in drawcto, in: Creativity and Cognition, ACM. p. 102–116. URL: http://dx.doi.org/10.1145/3635636.3656202, doi:10.1145/3635636.3656202.

Deshpande, M., Trajkova, M., Knowlton, A., Magerko, B., 2023. Observable creative sense-making (ocsm): A method for quantifying improvisational co-creative interaction, in: Creativity and Cognition, ACM. p. 103–115. URL: http://dx.doi.org/10.1145/3591196.3593514, doi:10.1145/3591196.3593514.

Ellard, K.K., Farchione, T.J., Barlow, D.H., 2011. Relative effectiveness of emotion induction procedures and the role of personal relevance in a clinical sample: A comparison of film, images, and music. Journal of Psychopathology and Behavioral Assessment 34, 232–243. URL: http://dx.doi.org/10.1007/s10862-011-9271-4, doi:10.1007/s10862-011-9271-4.

Fan, L., Zhuang, K., Wang, X., Zhang, J., Liu, C., Gu, J., Qiu, J., 2022. Exploring the behavioral and neural correlates of semantic distance in creative writing. Psychophysiology 60. URL: http://dx.doi.org/10.1111/psyp.14239, doi:10.1111/psyp.14239.

Forgeard, M.J., 2011. Happy people thrive on adversity: Pre-existing mood moderates the effect of emotion inductions on creative thinking. Personality and Individual Differences 51, 904–909. URL: http://dx.doi.org/10.1016/j.paid.2011.07.015, doi:10.1016/j.paid.2011.07.015.

Franceschelli, G., Musolesi, M., 2025. On the creativity of large language models. AI & society 40, 3785–3795.

Frazier, M.L., Johnson, P.D., Fainshmidt, S., 2013. Development and validation of a propensity to trust scale. Journal of Trust Research 3, 76–97.

Fukumura, K., Ito, T., 2025. Can llm-powered multi-agent systems augment human creativity? evidence from brainstorming tasks, in: Proceedings of the ACM Collective Intelligence Conference, ACM. p. 20–29. URL: http://dx.doi.org/10.1145/3715928.3737479, doi:10.1145/3715928.3737479.

Geroimenko, V., 2026. Collaborative and Co-creative Communication with AI. Springer Nature Switzerland, Cham. pp. 75–87. URL: https://doi.org/10.1007/978-3-032-21689-2_6, doi:10.1007/978-3-032-21689-2_6.

Glaveanu, V., Lubart, T., Bonnardel, N., Botella, M., Biaisi, P.M.d., Desainte-Catherine, M., Georgsdottir, A., Guillou, K., Kurtag, G., Mouchiroud, C., Storme, M., Wojtczuk, A., Zenasni, F., 2013. Creativity as action: findings from five creative domains. Frontiers in Psychology 4. URL: http://dx.doi.org/10.3389/fpsyg.2013.00176, doi:10.3389/fpsyg.2013.00176.

Green, A.E., 2016. Creativity, within reason: Semantic distance and dynamic state creativity in relational thinking and reasoning. Current Directions in Psychological Science 25, 28–35. URL: http://dx.doi.org/10.1177/0963721415618485, doi:10.1177/0963721415618485.

Hanchett Hanson, M., 2015. The ideology of creativity and challenges of participation. Europe's Journal of Psychology 11, 369–378. URL: http://dx.doi.org/10.5964/ejop.v11i3.1032, doi:10.5964/ejop.v11i3.1032.

Hearst, M.A., Allen, J., Guinn, C., Horvitz, E., 1999. Mixed-initiative interaction: Trends and controversies. IEEE Intelligent Systems 14, 14–23.

Heinen, D.J.P., Johnson, D.R., 2018. Semantic distance: An automated measure of creativity that is novel and appropriate. Psychology of Aesthetics, Creativity, and the Arts 12, 144–156. URL: http://dx.doi.org/10.1037/aca0000125, doi:10.1037/aca0000125.

Higgins, J.M., 1994. Creating creativity. Training & Development 48, 11–16.

Hubert, K.F., Awa, K.N., Zabelina, D.L., 2024. The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks. Scientific Reports 14. URL: http://dx.doi.org/10.1038/s41598-024-53303-w, doi:10.1038/s41598-024-53303-w.

Kaufman, J.C., Sternberg, R.J., 2010. The Cambridge handbook of creativity. Cambridge University Press.

Kim, K., Jeong, E., Lee, S., Yoon, S., Choi, Y.S., 2025. Claws:creativity detection for llm-generated solutions using attention window of sections, in: Belgrave, D., Zhang, C., Lin, H., Pascanu, R., Koniusz, P., Ghassemi, M., Chen, N. (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc.. pp. 133483–133516. URL: https://proceedings.neurips.cc/paper_files/paper/2025/file/c173f0a085728061bcd5b8a419170199-Paper-Conference.pdf.

Kim, S.J., 2024. Generative artificial intelligence in collaborative ideation: Educational insight from fashion students. IEEE Access 12, 49261–49274. doi:10.1109/ACCESS.2024.3382194.

Kumar, H., Vincentius, J., Jordan, E., Anderson, A., 2025. Human creativity in the age of llms: Randomized experiments on divergent and convergent thinking, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM. p. 1–18. URL: http://dx.doi.org/10.1145/3706598.3714198, doi:10.1145/3706598.3714198.

Li, Z., Liang, C., Peng, J., Yin, M., 2024. The value, benefits, and concerns of generative ai-powered assistance in writing, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, ACM. p. 1–25. URL: http://dx.doi.org/10.1145/3613904.3642625, doi:10.1145/3613904.3642625.

Lin, X., Huang, H., Huang, K., Shu, X., Vines, J., 2025. Seeking inspiration through human-llm interaction, in: Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, ACM. p. 1–17. URL: http://dx.doi.org/10.1145/3706598.3713259, doi:10.1145/3706598.3713259.

Lu, K., Qiao, X., Hao, N., 2019. Praising or keeping silent on partner's ideas: Leading brainstorming in particular ways. Neuropsychologia 124, 19–30. URL: http://dx.doi.org/10.1016/j.neuropsychologia.2019.01.004, doi:10.1016/j.neuropsychologia.2019.01.004.

Mersha, M.A., Kalita, J., 2025. Semantic-driven topic modeling for analyzing creativity in virtual brainstorming. URL: https://arxiv.org/abs/2509.16835, doi:10.48550/ARXIV.2509.16835.

Montag, C., Kraus, J., Baumann, M., Rozgonjuk, D., 2023. The propensity to trust in (automated) technology mediates the links between technology self-efficacy and fear and acceptance of artificial intelligence. Computers in Human Behavior Reports 11, 100315.

Nomura, M., Ito, T., Ding, S., 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, ACM. p. 1–9. URL: http://dx.doi.org/10.1145/3643562.3672609, doi:10.1145/3643562.3672609.

Pascual, D., Mur Dueñas, P., 2022. Dialogic interaction with diversified audiences in twitter for research dissemination purposes. Círculo de Lingüística Aplicada a la Comunicación 90, 61–79. doi:10.5209/clac.81307.

Pezzotti, N., Thijssen, J., Mordvintsev, A., Hollt, T., Van Lew, B., Lelieveldt, B.P., Eisemann, E., Vilanova, A., 2020. Gpgpu linear complexity t-sne optimization. IEEE Transactions on Visualization and Computer Graphics 26, 1172–1181. URL: http://dx.doi.org/10.1109/TVCG.2019.2934307, doi:10.1109/tvcg.2019.2934307.

Rammstedt, B., Kemper, C.J., Klein, M.C., Beierlein, C., Kovaleva, A., et al., 2013. A short scale for assessing the big five dimensions of personality: 10 item big five inventory (bfi-10). methods, data, analyses 7, 233–249.

Rouse, E.D., 2020. Where you end and i begin: Understanding intimate co-creation. Academy of Management Review 45, 181–204. URL: http://dx.doi.org/10.5465/amr.2016.0388, doi:10.5465/amr.2016.0388.

Ruangtanusak, S., Taveekitworachai, P., Pipatanakul, K., 2025. Talk less, call right: Enhancing role-play llm agents with automatic prompt optimization and role prompting. URL: https://arxiv.org/abs/2509.00482, doi:10.48550/ARXIV.2509.00482.

Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50, 696–735. URL: http://dx.doi.org/10.2307/412243, doi:10.2307/412243.

Sawyer, K., 2017. Group genius. 2 ed., Basic Books, London, England.

Sawyer, R.K., 2021. The iterative and improvisational nature of the creative process. Journal of Creativity 31, 100002. URL: http://dx.doi.org/10.1016/j.yjoc.2021.100002, doi:10.1016/j.yjoc.2021.100002.

Schock, L., Dyck, M., Demenescu, L.R., Edgar, J.C., Hertrich, I., Sturm, W., Mathiak, K., 2012. Mood modulates auditory laterality of hemodynamic mismatch responses during dichotic listening. PLoS ONE 7, e31936. URL: http://dx.doi.org/10.1371/journal.pone.0031936, doi:10.1371/journal.pone.0031936.

Seppänen, S., Makkonene, T., Tiippana, K., 2025. The psychophysiology of 'yes, and' vs. 'yes, but': The effect of acceptance, rejection and repetition during improvised dialogue. PsyArXiv doi:10.31234/osf.io/fa9dh_v1.

Shaer, O., Cooper, A., Mokryn, O., Kun, A.L., Ben Shoshan, H., 2024. Ai-augmented brainwriting: Investigating the use of llms in group ideation, in: Proceedings of the CHI Conference on Human Factors in Computing Systems, ACM. p. 1–17. URL: http://dx.doi.org/10.1145/3613904.3642414, doi:10.1145/3613904.3642414.

Singh, A.R., Netra, S., Vijarnia, A., Bhambri, S., Jain, A., 2025. A systematic framework for evaluating transformer architectures in semantic sentence similarity, in: 2025 IEEE International Conference on Contemporary Computing and Communications (InC4), IEEE. p. 1–6. URL: http://dx.doi.org/10.1109/InC465408.2025.11256279, doi:10.1109/inc465408.2025.11256279.

Smuts, H.E., 1992. An interactional approach to creativity. South African Journal of Psychology 22, 44–51. URL: http://dx.doi.org/10.1177/008124639202200202, doi:10.1177/008124639202200202.

Tang, L., 2025. The lack of other minds as the lack of coherence in human–ai interactions. Philosophies 10, 77. URL: http://dx.doi.org/10.3390/philosophies10040077, doi:10.3390/philosophies10040077.

Torrance, E.P., 1974. Torrance tests of creative thinking. Educational and psychological measurement .

Vaccaro, M., Almaatouq, A., Malone, T., 2024. When combinations of humans and ai are useful: A systematic review and meta-analysis. Nature Human Behaviour 8, 2293–2303. URL: http://dx.doi.org/10.1038/s41562-024-02024-1, doi:10.1038/s41562-024-02024-1.

Wagner III, J.A., 1995. Studies of individualism-collectivism: Effects on cooperation in groups. Academy of Management journal 38, 152–173.

Wang, Y., 2009. On cognitive foundations of creativity and the cognitive process of creation. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI) 3, 1–18.

Wei, B., Qiao, Y., Yin, J., 2025. Beyond the efficiency-meaning paradox: A collaborative flourishing framework for human-ai creative partnership URL: http://dx.doi.org/10.2139/ssrn.5855385, doi:10.2139/ssrn.5855385.

Ye, L., Sun, H., Zhang, J., Dong, B., Chu, X., Tao, J., Zhang, N., Zheng, X., Gong, R., 2024. Affect under need satisfaction and need thwarting: A new classification for the prediction of creative performance. Heliyon 10, e31323. URL: http://dx.doi.org/10.1016/j.heliyon.2024.e31323, doi:10.1016/j.heliyon.2024.e31323.

Yoo, J., Kim, J., 2013. Can online discussion participation predict group project performance? investigating the roles of linguistic features and participation patterns. International Journal of Artificial Intelligence in Education 24, 8–32. URL: http://dx.doi.org/10.1007/s40593-013-0010-8, doi:10.1007/s40593-013-0010-8.

Yousefi, M., Shahi, A., Sharifi, M., J Jorge Romera, A., Hoermann, S., Piumsomboon, T., 2025. Team dynamics in human-ai collaboration: Effects on confidence, satisfaction, and accountability, in: Proceedings of the 27th International Conference on Multimodal Interaction, ACM. p. 395–404. URL: http://dx.doi.org/10.1145/3716553.3750776, doi:10.1145/3716553.3750776.

Zhou, Z., Li, J., Zhang, Z., Yu, J., Duh, H., 2024. Examining how the large language models impact the conceptual design with human designers: A comparative case study. International Journal of Human–Computer Interaction 41, 5864–5880. URL: http://dx.doi.org/10.1080/10447318.2024.2370635, doi:10.1080/10447318.2024.2370635.

Creativity Is Not an Outcome: A Process-Based Framework for Mapping Individual Creative Contributions in Human–LLM Collaboration

Nishthaa Lekhi[a], Trevor Patten[c], Prakash Patil[b], Mengyao Li[c], Areen Alsaid[b,*]

[a] Human-Centered Design Engineering, University of Michigan–Dearborn, 4901 Evergreen Road, Dearborn, MI 48128, USA

[b] Department of Industrial & Manufacturing Systems Engineering, University of Michigan–Dearborn, 4901 Evergreen Road, Dearborn, MI 48128, USA

[c] School of Psychology, Georgia Institute of Technology, North Avenue, Atlanta, GA 30332, USA

*Corresponding author: Areen Alsaid
Department of Industrial & Manufacturing Systems Engineering
University of Michigan–Dearborn
4901 Evergreen Road, Dearborn, MI 48128, USA
Email: alsaid@umich.edu

Email addresses:
nlekhi@umich.edu (N. Lekhi);
tpatten7@gatech.edu (T. Patten);
patilpb@umich.edu (P. Patil);
mengyao.li@gatech.edu (M. Li);
alsaid@umich.edu (A. Alsaid)