Playing Along:
Building AI Agents for Co-Creation of Improvised Stories

Idan Dov Vidra[1] and **Gal Kimron**[1] and **Lior Noy**[2] and **Ariel Shamir**[1]
[1]Reichman University, arik@runi.ac.il, idan.vidra@post.runi.ac.il, gal.kimron@post.runi.ac.il
[2]Ono Academic College, lior.noy@ono.ac.il

Abstract

This paper studies human-agent co-creation of improvised stories, investigating whether Large Language Models can effectively engage in an improvisational practice known as the “Yes! and...” game. We demonstrate how AI systems can participate successfully in improvisational co-creation, moving beyond response generation to collaborative story-telling. We provide a systematic framework for evaluating creative AI outputs in improvisational contexts, combining human evaluation with computational metrics. Our evaluations show that stories co-created with an AI agent received similar ratings to human-human collaborations and were perceptually indistinguishable in blind testing. This shows how AI’s creative invention abilities - which can be limitations in factual tasks - become assets in collaborative storytelling. More generally, our approach presents the “Yes! and...” game as a novel model-system for studying improvised co-creativity in a well-defined and measurable setup.

Introduction

While co-creative storytelling and human-AI collaboration received growing attention (Rezwana and Maher 2023; Branch, Mirowski, and Mathewson 2021; Concepcion, Gervas, and Mendez 2019; Veale, Wicke, and Mildner 2019; Jacob et al. 2019; Branch et al. 2024), creative improvisation, particularly in a minimal, text-based setting, remains underexplored. This work advances computational creativity in three ways. First, it demonstrates how AI systems can participate successfully in improvisational co-creation, moving beyond response generation to collaborative storytelling. Second, it provides a systematic framework for evaluating creative AI outputs in improvisational contexts, combining human evaluation with computational metrics. Third, it shows how AI’s narrative invention abilities serve as creative assets in collaborative storytelling.

We focus on human-AI co-creation of improvised stories, investigating whether Large Language Models (LLMs) can effectively engage in a specific improvisational practice - the “Yes! and...” game. The “Yes! and...” game applied here is an example of a fundamental tenet of improvisation, the “Yes! and...” approach (Berliner 2009; Nachmanovitch 1991; Johnstone 1987), where improvisers need both to accept their partners’ offers (the *Yes!* part) and then continue the flow from this position (the *and...* part). In the “Yes! and...” game we apply here, two participants build a shared ”adventure” by always accepting and elaborating on each other’s contributions (Johnstone 1987).

The “Yes! and...” approach is different from other forms of narrative co-creation. Unlike turn-based storytelling or guided narrative generation, participants must respond spontaneously without planning while building upon each other’s contributions. This creates a balanced creative dynamic of acceptation and extension, where both participants actively shape the emerging narrative through real-time interaction.

The narratives created are jointly invented and participants inevitably generate *imaginative* content. For an AI agent to participate effectively, it must therefore engage in creative invention — envisioning shared experiences and fictional scenarios that go beyond factual content.

LLMs, trained on vast text corpora, can produce coherent documents and engage in conversations (Shanahan 2023; Zhao et al. 2023; Liu et al. 2023; Brown et al. 2020; Thoppilan et al. 2022). While these models often produce hallucinations— ungrounded outputs that deviate from factual accuracy (Agrawal, Mackey, and Kalai 2023; Ji et al. 2023)— such outputs are increasingly seen as valuable in creative processes (Jiang et al. 2024; Halperin and Lukin 2024; Chakraborty and Masud 2024). This aligns with foundational theories of AI creativity: Boden’s categorization suggests that “imagined” elements expand narrative search spaces, while Ritchie’s criteria emphasize balancing novelty and usefulness (Boden 2004; Ritchie 2001). Established frameworks like SPECS (Jordanous 2012) and the Novelty-Value-Surprise model (Grace et al. 2015) ground our approach to evaluating creativity in co-created stories.

With the “Yes! and...” game we adopt a reduced, model-system approach: we use a textual “Yes! and...” scenario that strips away visual, auditory, and embodied dimensions. Similar to model systems in other experimental paradigms (Noy, Dekel, and Alon 2011; Noy 2014; Rafner et al. 2023), this minimalistic setup isolates core mechanisms of co-creative improvisation and enables systematic analysis. To investigate whether LLMs can participate in “Yes! and...” games at a level comparable to humans, we developed AI agents using a chat interface and evaluated the resulting stories through human ratings, computational metrics, and a “Turing Test”-style survey.

Our main contributions include introducing a text-only version of the "Yes! and..." game for human-human (HH) interactions and extending it to human-AI (AIH) partnerships, demonstrating through empirical evaluation that LLM-based agents achieve comparable ratings to humans across creativity metrics and are indistinguishable in blind testing. We develop and validate evaluation measures combining human ratings with computational metrics of novelty and surprise, showing strong inter-rater reliability ($ICC > 0.75$) on the two core creative dimensions—creativity and surprise—while reliability for the remaining dimensions was moderate to poor. Our minimalistic paradigm enables controlled study of human-AI creative collaboration, building on model system approaches from joint action research (Noy, Dekel, and Alon 2011; Noy 2014) to isolate core mechanisms of improvisational interaction.

Previous Work

Our work builds upon three main research areas: computational co-creativity, improvisational AI, and creative evaluation frameworks.

Computational Co-Creativity with LLMs Recent work has demonstrated LLMs' potential for creative collaboration with humans (Franceschelli and Musolesi 2024; Yuan et al. 2022; Gero, Liu, and Chilton 2022). Although traditionally viewed as flaws for factual tasks, LLMs' creative inventions are increasingly recognized as valuable in creative contexts (Jiang et al. 2024; Halperin and Lukin 2024; Chakraborty and Masud 2024). Our work extends this perspective by showing how apparent limitations can become assets in improvisational storytelling.

Improvisational AI Systems Early pioneering work by Hayes-Roth on directed improvisation (Hayes-Roth et al. 1994) and Bates' Oz Project (Bates, Loyall, and Reilly 1994) established foundational principles for computational improvisation. Building on these foundations, recent research has focused primarily on music (Hoffman and Weinberg 2010) and theatrical performance (Branch et al. 2024). Researchers have explored various aspects of improvised interaction, from movement-based object interactions (Jacob et al. 2019) to complementary AI partnerships in storytelling (Veale, Wicke, and Mildner 2019). Recent work has extended this to evaluating LLMs in professional theatrical settings (Branch et al. 2024). Particularly relevant is work by (Cho and May 2020) on dialogue systems adhering to improvisational principles. We build upon these foundations while introducing a minimalist text-based framework that enables systematic study of human-AI creative dynamics.

Evaluation Frameworks Our evaluation approach draws on established frameworks for assessing computational creativity, particularly SPECS (Jordanous 2012) and the Novelty-Value-Surprise model (Grace et al. 2015). We extend these frameworks by combining human evaluation with computational metrics specifically designed for improvisational co-creation. This extends previous work on evaluating co-creative systems while introducing new measures for improvisational interaction (Karimi et al. 2018).

Our approach advances existing work in three key ways: (1) we introduce a minimal, controlled environment for studying human-AI improvisation, (2) we provide evaluation metrics specifically designed for improvisational co-creation, and (3) while some recent work has begun to recognize the creative potential of LLM beyond factual constraints (Jiang et al. 2024; Halperin and Lukin 2024), we demonstrate how these generative capabilities become creative assets in improvisational storytelling through empirical evaluation.

Methods

Textual "Yes! and..." Game for Human Interaction

We developed a chat-based platform that allows pairs of participants to engage in "Yes! and..." games through text-based interaction. We call these human-human (HH) interactions.

We randomly paired participants and instructed them to play an improvisational joint storytelling game following the "Yes! and..." principle. After reading the instructions, participants took turns creating sentences that evolved their shared story. The first participant was chosen at random. While we suggested starting each contribution with "Yes! and...", this was not enforced. We placed no restrictions on sentence length, resulting in an average of 11.35 words per sentence and 13.9 sentences per game.

Game Play The game sessions followed a structured progression: participants first played a short game, followed by a longer game (240-360 seconds), and finally an unrestricted game where they could reach a natural conclusion. Some participants engaged in this sequence with multiple partners, returning to the platform's main window to be randomly paired again. For privacy protection, we only recorded the text of the story and the timestamps, without collecting personal information.

Following each session, participants participated in a structured debriefing in which they could share their experiences and learn about the research goals. Additional details about game protocols, data collection, and full game examples are provided in the Supplemental Information (SI).

AI-Human "Yes! and..." Game Implementation

We developed a digital platform to enable "Yes! and..." games between human participants and an agent based on GPT-4. This platform was similar (but not identical) to the one used for HH interactions (see SI: Interface Screenshots). The primary difference was the incorporation of GPT-4 as one of the two players. The Gradio package (Python) was utilized for managing the chat interactions (Abid et al. 2019). We call these AIH-interactions. For technical reasons, all AIH games were designed so that humans made the first contribution to the story.

Game Play Participants were initially led to believe they were interacting with another human. Participants received exactly the same instructions as those in the HH game sessions. This approach was adopted to study natural interaction without preconceived biases from playing with an agent. After each session, participants were informed that their counterpart was actually an AI model.

Model Configuration The agent used GPT-4 accessed through OpenAI's Chat Completion API, configured with a presence penalty of 1.7 to encourage response diversity while maintaining default temperature. To mirror human interaction patterns, we introduced random delay between 21-25 seconds in the model responses, based on an analysis of HH response times, we also limited responses to 11 words. Games were designed to conclude after approximately 10 messages, either through participant choice or an AI-generated concluding message. Each interaction incorporated the full conversation history for context coherence.

Model's Prompt We refined the prompt used for GPT-4 based on participants' feedback, avoiding responses which were too quick or too long, and seemed non-human (see complete list of suggestions in SI: Development of the AI Agent's Prompt). The final prompt was:

> "You are a model built to play an interactive "Yes! and..." game with people. In this game, we'll build a story together by taking turns adding sentences. Each addition should build upon the previous statement in the context of the entire story, showing agreement and creativity. Be adaptable and willing to explore various tones, including darker themes. Feel free to start with "Yes! and..." or other affirming phrases like 'Yeah...', 'Totally...', 'Right and...', 'definitely...'. Keep responses concise, engaging, and in simple language, using 11 words or less."

AIH games typically contained up to six sentences from the GPT-4 model, similar to the average length of HH games. These constraints resulted in interactions that closely matched the natural flow of HH games while maintaining consistent engagement throughout the session.

Data Collection and Analysis

We collected stories through various platforms, gathering 141 HH games from 82 participants and 86 AIH games from 45 participants (collection methods detailed in SI: Summary of HH Game Collection Methods, Statistics for AIH Games Collections). Story lengths varied by design, as participants progressed from short to longer to unrestricted games. All stories underwent quality filtering based on two criteria: minimum length ($>$ 3 lines) and structural coherence (adherence to "Yes! and..." game structure and basic narrative flow). After filtering, our final dataset comprised 129 HH and 75 AIH stories. Estimated game durations, based on average response times per turn, averaged 548.71 seconds (SD = 220.82), with the high variance reflecting our intentionally varied game durations (detailed timing analysis in SI: Estimated Game Times).

Human-Human Stories Collection methods evolved from initial Google Docs to a dedicated web platform. Stories were excluded for containing fewer than three lines (5 stories) or failing quality criteria (10 stories) as judged independently by two researchers. The final HH dataset averaged 18.42 sentences (SD = 4.28) and 8.81 words per sentence (SD = 2.21).

AI-Human Stories AIH stories were collected using GPT-4 as the AI agent, selected based on preliminary tests demonstrating its superior engagement compared to other models. Data collection evolved through three phases, with progressive refinements to the prompt and interface based on participant feedback (details in SI: Statistics for AIH Games Collections). The system concatenated all previous messages as context for each AI response. The final AIH dataset averaged 13.08 sentences (SD = 5.3) and 14.13 words per sentence (SD = 4.6).

Story Evaluation Process

A total of 453 participants from university took part in the experiment, without prior experience in improvised story co-creations. We recruited participants from the student body, with the incentive of receiving course credit equivalent to 30 minutes of their time. The experiment was approved by the ethical committee and was performed on Qualtrics.

To reduce the total evaluation time per participant, we randomly selected 40 HH stories and 24 AIH stories from our dataset, excluding stories that deviated from the narrative-building format or contained excessive context-specific references (selection criteria detailed in SI: Data Selection and Preprocessing). All non-English stories were translated to English before evaluation. We divided these 64 stories into eight batches, each participant evaluating one batch of 8 stories presented in random order. Initial cohort sizes were 68, 64, 98, 100, 97, 85, 145, and 164 participants, respectively, although these numbers were later reduced through our filtering process (see Filtering of Evaluators below).

We first introduced participants to the concept of the "Yes! and..." game via an introductory video. Participants then evaluated each story on five criteria (creativity, interest, surprise, cohesiveness, and degree of agreement) using a 7-point Likert scale (1=lowest, 7=highest).

Following the evaluations, we collected participants' experiences through a structured debriefing. Participants also self-assessed their English proficiency on a scale of 1-7, with most (84.3%) rating themselves at level 5 or higher, and indicated whether they had previously evaluated "Yes! and..." games. The survey concluded with an opportunity for open-ended feedback. Additional details about survey design, participant demographics, and response statistics are available in the SI.

Rationale and Details of Evaluation Criteria

A co-created "Yes! and..." story, composed of 10-20 sentences, is a multifaceted object that requires careful evaluation. Through open-ended interviews with three improvisation experts who evaluated 15 HH stories, we formulated five evaluation criteria specific to "Yes! and..." stories:

Creativity Creativity is pivotal in “Yes! and...” games because of its role in pushing narrative boundaries. It is not just about being different; it is about generating narratives that promote divergent thinking and challenge the conventional patterns of storytelling (Sowden et al. 2015).

Interest The ability to engage the audience is fundamental in these games. Although structured narrative principles help maintain engagement (McKee 1997), improvisational storytelling can create compelling narratives through character interaction even without predetermined plots (Swartjes and Vromen 2007).

Surprise In “Yes! and...” games, surprise is a key measure due to its role in maintaining audience engagement through unexpected narrative shifts. This concept aligns with (Ely, Frankel, and Kamenica 2015) findings on the importance of surprise in the entertainment value of information.

Cohesiveness Cohesiveness in “Yes! and...” games helps to maintain a structured and understandable narrative, balancing spontaneous creation with causal connection between the elements of the story (Alon 2010; McKee 1997). Emergent narrative studies show that while interactive freedom can challenge coherence, successful improvisation maintains the essential narrative structure through collaborative effort (Theune et al. 2013).

Degree of Agreement Between Players This criterion assesses how well participants align in their vision for the story’s development, following Johnstone’s principle that successful improvisation requires accepting and building upon each other’s offers (Johnstone 1987). This mutual acceptance is the foundational “Yes” in the “Yes! and...” principle, enabling collaborative story development.

The complete survey definitions and rating instructions for each criterion are provided in SI: Section Evaluation Criteria Details.

Filtering of Evaluators

In our study, we implemented a filtering process to ensure the participation of the evaluator using three criteria:

First, we excluded raters who took over 1000 seconds (approximately 17 minutes) to complete the survey of eight stories, as this duration was substantially longer than the average completion time of 321.35 seconds (±38.65s) and suggested potential disengagement or interruptions during the task. This eliminated 49 of 453 raters.

Second, for each story, we calculated raters’ average time and SD for answering questions. Raters who spent less than (mean - 0.85 * SD) were considered insufficiently engaged and excluded from all games in the survey. This method, while potentially excluding some accurate raters, prioritized data quality by ensuring thoughtful participation. Across the eight cohorts, we removed 5-9 raters per cohort for taking too long (totaling 49 raters) and 0-23 raters per cohort for rating too fast (totaling 85 raters).

Finally, we filtered out participants based on their self-reported English fluency. Specifically, those who rated their proficiency as “Low” or “Not Good” (the bottom two levels out of a seven-tier scale) were excluded. This resulted in removing 11-27 participants per cohort (totaling 151 participants).

Following the application of these filters, the remaining cohort sizes were 45, 47, 61, 56, 44, 34, 116 and 133 participants respectively.

Computational Metrics: Novelty, Surprise, and Value

Our evaluation approach combines traditional human evaluation with computational metrics, reflecting recent advances in creativity measurement. Although classic creativity tests such as the Alternative Uses Test relied solely on human assessors to assess creative output (Kudrowitz and Dippo 2013), modern approaches can take advantage of LLMs to provide an automated assessment of creative qualities like originality and quality (Luchini et al. 2023). Following this trajectory, we combine human ratings with automated metrics to provide both subjective evaluation and objective measurement of the characteristics of the story.

We employed three metrics to assess creative aspects of the co-created “Yes! and...” stories: *Novelty*, *Surprise*, and *Value (Engagement)*. These metrics were chosen specifically to capture key aspects of successful “Yes! and...” interactions - the ability to build upon previous contributions while introducing new elements (Novelty), the capacity to take unexpected narrative turns (Surprise), while maintaining coherence (evaluated through cohesiveness), and the sustained engagement of participants (Value). Novelty was computed via established embedding-based approaches (Johnson et al. 2022; Julian Just and Hutter 2024), measuring semantic distances between consecutive contributions, while Surprise was quantified using language model probability distributions (Bunescu and Uduehi 2022). Value, unlike these computational metrics, was derived primarily from participant behavior and feedback.

Novelty (Semantic Variability) In “Yes! and...” games, participants must balance between maintaining narrative coherence and introducing new ideas. To measure this balance quantitatively, we embedded each sentence using pre-trained transformer models (e.g., `roberta-base-nli-stsb-mean-tokens` (Liu et al. 2019)). For each story, we computed the cosine distance between each sentence and the sentence immediately preceding it (excluding the first). Higher mean distances indicate stronger semantic shifts between consecutive contributions, suggesting more novel narrative directions. We analyzed these distances separately for human-authored lines in both conditions (HH and AIH) and AI-authored lines in the AIH condition.

Surprise (Unpredictability) We define Surprise as the negative log probability that a large language model assigns to each new sentence, given its immediate context. Concretely, we use pretrained models such as `GPT-2` or `OPT-1.3b` to estimate:

$$\text{Surprise} = -\log\Big(P(\textit{sentence} \mid \textit{previous sentences})\Big)$$

Thus, sentences that are unlikely or dissimilar to the model's learned language patterns yield higher Surprise scores. We selected these specific models due to practical computational constraints and their widespread availability through the HuggingFace library. We computed Surprise separately for human-authored lines and AI-authored lines in the AIH condition, as well as for all lines in HH stories.

Value (Engagement) Unlike Novelty and Surprise, which rely on embedding and probability metrics, Value focuses on participants' *engagement*. We operationalized Value in three ways:

1. **Session Duration and Turns:** The time span of each game (in seconds) and the number of sentences exchanged. Longer or more extensive interactions suggest higher engagement.
2. **Participant Feedback:** Open-ended comments solicited immediately after each session. We looked for indications of immersion, boredom, or heightened excitement.
3. **Reported Interest Ratings:** As part of the external evaluation, naive raters judged how *interesting* each story was, on a scale from 1 (least) to 7 (most). We interpret higher interest ratings as a proxy for story Value.

Ablation Study: Prompt without "Surprise"

To examine how explicitly prompting the AI to be "surprising" affects its output, we conducted an ablation in which the word "surprise" was removed from the prompt. The rest of the instructions were unchanged (see the previous discussion for the baseline prompt details). The participants again participated in AIH games, but this time the AI instructions did not refer to "surprising" content. We collected the resulting 30 stories and recalculated the Surprise metric to see whether removing the explicit instruction to be 'surprising' would alter the creative dynamics.

"Turing Test" for HH and AIH Games

We conducted an evaluation to compare narratives generated by HH pairs and AIH pairs in "Yes! and..." games. This evaluation, framed as a "Turing test", aimed to explore the capability of AI models in computational creativity through the lens of co-created narratives. 150 university students participated in the survey, which included 10 randomly selected stories (5 HH and 5 AIH). We asked participants to identify whether each story was created by HH or AIH pairs, with an optional explanation field (see survey interface in SI Turing Test Survey Interface and Implementation). After filtering for English proficiency, 122 raters remained.

Results

Stories Data

We analyzed 141 HH and 86 AIH games collected through various platforms. After filtering for quality and minimum length requirements (detailed filtering criteria in Methods - Filtering of Evaluators), our final dataset comprised 129 HH and 75 AIH stories. Games averaged 13-18 sentences in length.

Table 1: Comparison of length of interactions between humans and AI agents, in HH and AIH corpora (means and SD).

Metric	AI	H in AIH	H in HH
Sentences per story	6.52 (2.64)	6.52 (2.63)	6.95 (2.32)
Words per sentence	16.66 (7.59)	11.51 (3.58)	11.37 (3.17)
Words per story	109.05 (64.64)	77.37 (41.86)	79.07 (31.77)

Below are partial examples of the interactions in both conditions (full examples of both HH and AIH games are provided in SI: Game Examples):

HH:

Remember the other day we saw flying whales?

Yes of course! One of them offered us a ride to Paris and we got on.

Yes, and then at the end we got off one stop before, on the moon.

AIH:

Do you remember that day at the park?

Yes, and especially how the golden autumn leaves danced in the wind.

Yes! And that wind! It was so powerful!

In the AIH examples, human text is shown in blue and AI text in green.

All HH and AIH interactions used in this study are available in our public corpus repository: `https://tinyurl.com/yes-and-github`.

Statistics of Co-Created Stories

We analyzed basic linguistic metrics (sentences per story, words per sentence and total words) to compare human and AI behaviors in these co-created stories. Table 1 presents the means and standard deviations for each metric across conditions.

Analysis of variance (ANOVA) revealed that while the number of sentences per story was similar across conditions, AI agents produced significantly more words per sentence than humans (mean difference ≈ 5.2 words, F(2, 407) = 43.70, $p < 0.001$) and consequently more words per story (mean difference ≈ 30 words, F(2, 407) = 16.29, $p < 0.001$). This verbose AI behavior did not influence human partners' response lengths, as humans maintained consistent word counts whether paired with AI or other humans (mean difference = 0.14 words, p = 0.952). Detailed statistical analyses, including post-hoc comparisons, are provided in SI: Detailed Statistical Analysis of Co-Created Stories.

Comparing Evaluations of HH and AIH Stories by External Raters

The collected "Yes! and..." stories were evaluated by 453 unique raters, with a total of 817 ratings (see details in SI: Survey Participation Statistics). In each rating session, a naive rater (who did not play the game before) read and rated 8 stories. Each story was evaluated on five criteria: creativity, interest, surprise, cohesiveness, and the degree of agreement between players, using a 7-point Likert scale (see details of these criteria in Methods).

To assess inter-rater reliability, we used the ICC3K variant of the intraclass correlation coefficient, which measures consistency among raters using a 7-point scale (Koo and Li 2016). Overall agreement was good (mean ICC = 0.761, SD = 0.154), with particularly strong reliability for creativity (ICC = 0.809) and surprise (ICC = 0.763) ratings. However, agreement (ICC = 0.215) and story cohesiveness (ICC = 0.465) showed poor reliability, while interest ratings were moderate (ICC = 0.544). Details in SI: Detailed Intraclass Correlation Coefficient (ICC) Analysis.

Evaluation Results of HH and AIH Stories

Table 2 presents external raters' evaluations across all criteria on a 7-point Likert scale. Statistical analysis revealed remarkably similar ratings between HH and AIH stories, with only the "interesting" criterion showing a marginal difference favoring AI-human pairs (3.72 vs 3.58, $p = 0.011$). Notably, both conditions received above-median ratings (> 3.5) for all criteria. Details in SI: Detailed Statistical Analysis of Story Evaluations.

Table 2: Mean and standard deviation (STD) of ratings for each criterion in HH and AIH "Yes! and..." games.

Criterion	Mean HH	STD HH	Mean AI	STD AI
Creative	4.26	1.88	4.34	1.78
Agreement	4.57	1.87	4.55	1.82
Story	4.26	1.83	4.25	1.77
Interesting	3.58	1.86	3.72	1.81
Surprising	3.83	1.96	3.87	1.89

A key finding of our analysis is the striking similarity between HH and AIH rating distributions, visualized using Jensen-Shannon Divergence in Figure 2. This information-theoretic measure (Lin 1991) quantifies the similarity between probability distributions, with darker shades indicating lower similarity. The analysis shows strong similarities across all evaluation criteria, supporting our statistical findings.

The similarity between HH and AIH stories is further supported by detailed distribution analysis in Figure 1 - allowing direct comparison between HH and AIH games. Analysis using only non-repeated raters (as described in Methods) yielded similar results (see SI: Analysis of Non-Repeated Raters).

Novelty, Surprise, and Value Analysis

Novelty Results To assess semantic variability in story progression, we compared consecutive-sentence distances across three groups: humans in human-human games (H_HH), humans in AI-human games (H_AIH), and AI responses in AI-human games (AI_AIH). Higher distances indicate greater semantic novelty between successive sentences. Analysis using multiple embedding models showed consistent patterns (Table 3):

Table 3: Mean consecutive-sentence distances by group and embedding model (with SD).

Model	H_HH	H_AIH	AI_AIH
RoBERTa	0.758 (0.182)	0.783 (0.176)	0.616 (0.184)
DistilBERT	0.763 (0.192)	0.771 (0.173)	0.624 (0.174)
MPNet	0.644 (0.167)	0.657 (0.145)	0.566 (0.152)
MiniLM	0.634 (0.182)	0.670 (0.163)	0.584 (0.153)

Statistical analysis revealed significant differences between AI-authored and human-authored contributions ($p < 0.001$ across all models) - with humans being more "novel". Notably, humans collaborating with AI (H_AIH) showed slightly higher semantic distances than those in purely human interactions (H_HH), though this difference was only significant for two of the four models ($p < 0.05$ for RoBERTa and MiniLM).

As visualized in Figure 3, humans consistently demonstrated greater semantic leaps between contributions compared to AI across all embedding models.

Surprise Results The analysis of sentence unpredictability using multiple language models revealed systematic differences between contributions. Higher scores indicate more unexpected sentences given the previous context.

Figure 4 illustrates these surprise patterns across participant groups and language models.

Table 4 summarizes these findings:

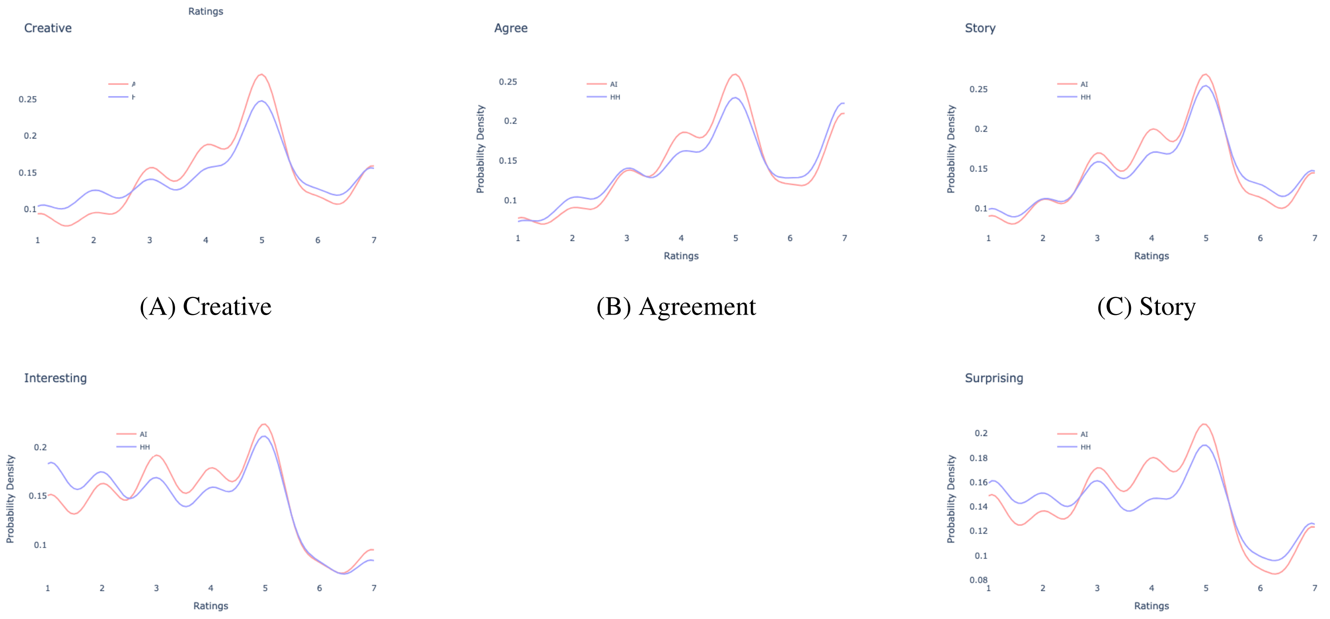
Table 4: Mean surprise scores by group and language model (with standard deviations).

Model	H_HH	H_AIH	AI_AIH
GPT-2	47.75 (24.92)	59.23 (27.19)	74.43 (29.80)
BLOOM-1B1	46.29 (23.69)	55.95 (25.02)	70.53 (27.85)
OPT-1.3B	44.15 (22.98)	52.87 (23.65)	64.62 (24.55)

All models showed a consistent pattern: AI-authored sentences (AI_AIH) exhibited significantly higher surprise scores than human-authored sentences in either condition ($p < 0.001$). Additionally, humans paired with AI (H_AIH) produced more surprising responses than those in human-human pairs (H_HH), suggesting that AI collaboration may encourage more unexpected narrative turns.

While both novelty and surprise measure aspects of unexpectedness, they capture different phenomena. Novelty measures semantic distance between consecutive contributions regardless of context (Figure 3), while surprise measures contextual unexpectedness given the specific preceding narrative (Figure 4). The fact that AI contributions show lower novelty (smaller semantic leaps) but higher surprise (more unexpected in context) suggests these metrics indeed capture distinct aspects of creative behavior.

Ablation Study Results To investigate the impact of explicitly prompting for surprise, we conducted an ablation study where surprise-related instructions were removed from the AI's prompt. Table 5 compares the mean surprise



(A) Creative

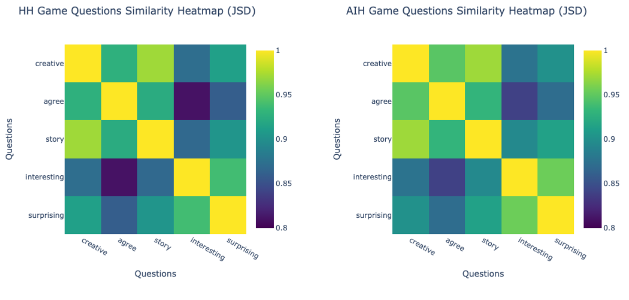
(B) Agreement

(C) Story

(D) Interesting

(E) Surprising

Figure 1: Complete rating distributions across all evaluation criteria. Each subplot shows the probability density distributions for both HH (blue) and AIH (red) games. The distributions demonstrate remarkable similarity between human-human and AI-human interactions.



(a) HH Games

(b) AIH Games

Figure 2: Rating distribution similarity heatmaps. Darker shades show lower similarity.

scores between the original and ablation conditions across both human and AI outputs.

Table 5: Mean surprise scores for original and ablation conditions (with standard deviations).

Condition	**Human**	**AI**
Original (AIH)	59.10 (27.36)	74.75 (30.18)
Ablation (ABL)	58.64 (20.04)	67.55 (12.35)

As shown in Figure 5, removing 'surprise' from the AI's instructions significantly affected AI outputs ($p = 0.002$, Cohen's $d = 0.312$), with mean surprise scores dropping from 74.75 to 67.55. However, human responses remained largely unchanged ($p = 0.703$, Cohen's $d = 0.019$), suggesting that explicit surprise prompting primarily influences AI behavior rather than human creativity. The gap between human and AI surprise scores remained substantial in both the original condition ($p < 0.001$, Cohen's $d = -0.543$) and the ablation condition ($p < 0.001$, Cohen's $d = -0.535$), indicating that AI responses maintain higher unpredictability regardless of explicit prompting - a notable pattern differing from other creative domains (Hoffman and Weinberg 2010).

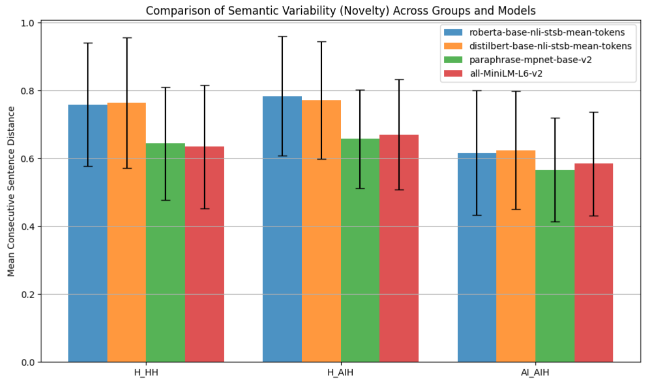


Figure 3: Novelty comparison: Semantic distances between consecutive sentences across different embedding models. Higher values indicate greater semantic novelty between successive contributions.

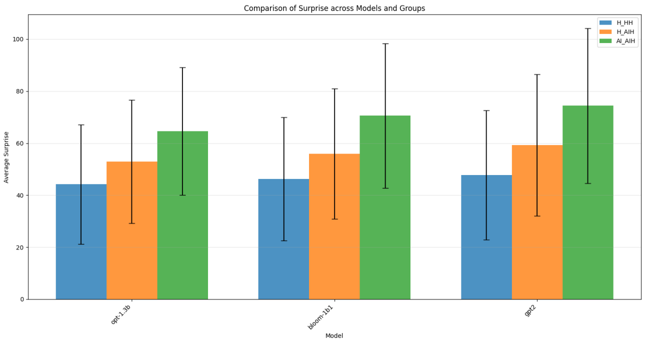


Figure 4: Surprise comparison: Scores across different language models. Higher values indicate more unexpected sentence constructions given the context.

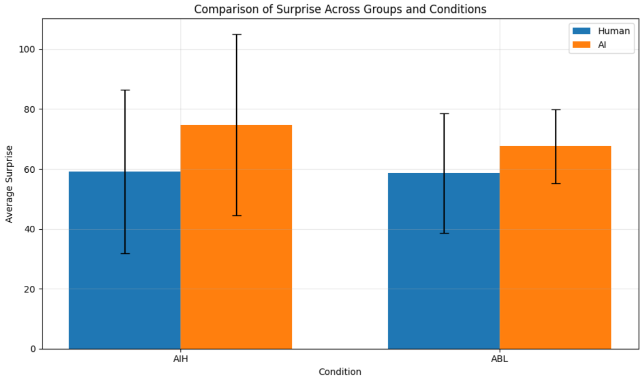


Figure 5: Ablation study results comparing surprise scores between original and modified conditions, demonstrating the effect of removing surprise-related prompting from AI instructions.

Value (Engagement) Analysis Analysis of value metrics through engagement revealed sustained participation across both HH and AIH conditions. Game durations remained comparable (HH: 508.63 ± 169.25 seconds; AIH: 405.73 ± 234.98 seconds), and sentence lengths were consistent between human players across formats (HH: 11.37 words; AIH: 11.51 words). External raters judged AIH stories as marginally more interesting than HH stories (3.72 vs. 3.58, $p = 0.011$), though this difference was modest compared to other evaluation criteria.

A simple 'Turing Test' for "Yes! and..." Stories

To assess whether raters could distinguish between HH and AIH "Yes! and..." stories, we conducted a survey involving 122 raters (see more details in Methods). Each rater was presented with 5 HH games and 5 AIH games (chosen randomly) and was asked to identify whether each story was created by a pair of humans or a human and AI. Overall accuracy was poor, with 47.05% correct identification for HH stories and 46.56% for AIH stories. A Chi-square test confirmed no significant difference in raters' ability to identify story sources ($\chi^2 = 0.0132$, p = 0.9086).

Table 6: Raters' guesses and Chi-square test results for each "Yes! and..." story, showing the number of correct and incorrect guesses, Chi-square test statistic, p-value, and whether the difference in guesses was significant.

Story	Correct	Incorrect	Chi-sq	P-val	Sig.
HH 1	40.16%	59.84%	8.67	0.003	Yes
HH 2	46.72%	53.28%	0.80	0.370	No
HH 3	55.74%	44.26%	2.77	0.096	No
HH 4	47.54%	52.46%	0.41	0.522	No
HH 5	45.08%	54.92%	1.98	0.159	No
AIH 1	36.07%	63.93%	17.85	< 0.001	Yes
AIH 2	47.54%	52.46%	0.41	0.522	No
AIH 3	48.36%	51.64%	0.15	0.701	No
AIH 4	51.64%	48.36%	0.15	0.701	No
AIH 5	49.18%	50.82%	0.02	0.898	No

Story-level analysis (Table 6) revealed that for both HH and AIH stories, in 4/5 cases there were no significant differences between correct and incorrect responses, and in 1/5 stories there were significantly more incorrect responses (more details in SI: Detailed Turing Test Results). Taken together, these results suggest that naive readers cannot differentiate between stories created in the HH and AIH conditions.

To gain insights into the factors that influenced raters' decisions when distinguishing between HH and AIH "Yes! and..." games, we collected testimonials from participants explaining their reasoning. These testimonials provide qualitative data that complement the quantitative results presented earlier.

The following are representative examples of the testimonials, illustrating the various factors that raters considered when making their judgments:

Explanations from People Who Mistakenly Thought a Human-Made Story Was Made by a Human and AI

> "The writing and context here are perfect. Definitely written with AI."

> "It seems to me that this story was written between a person and an AI [...] The story includes exaggerated fictional events that sound improbable in reality [...] The conversation sounds forced and lacks logical connection [...] There is not much emotion or human reactions in the conversation."

Explanations from People Who Mistakenly Believed a Story Made by a Human and AI Was Actually Made by Just Humans

> "AI would not write the following sentence ever: 'Yes and I never saw something like that in my life'."

> "There is a lot of text and detail, too much science fiction."

Discussion

Analysis of Results Our computational analysis revealed distinctive patterns in AI versus human contributions to the

"Yes! and..." game. AI responses showed higher surprise scores across all evaluation models (mean difference ≈ 15-20 points (see in Methods), $p < 0.001$) but lower semantic novelty compared to humans (mean difference ≈ 0.15 points, $p < 0.001$ across all embedding models). AI responses were consistently longer (mean difference 5.2 words per sentence, $p < 0.001$) but didn't influence human partner response lengths, suggesting humans maintained their natural interaction style regardless of partner type. External raters judged AIH stories as comparable to HH stories across all evaluation criteria except interest, where AIH stories scored marginally higher (3.72 vs 3.58, $p = 0.011$). Notably, in our "Turing test" experiment, participants were unable to reliably distinguish between HH and AIH stories (47.05% vs 46.56% correct identification), indicating that AI contributions were naturalistic enough to be indistinguishable from human responses.

Comparison with Literature These findings extend previous work on LLMs in creative writing (Franceschelli and Musolesi 2024) and storytelling (Yuan et al. 2022) by demonstrating their capacity for real-time improvisational co-creation. The "Yes! and..." principle, long recognized as a basic tenet of improvisation (Johnstone 1987) and collaborative creativity (Himberg et al. 2018), provides a structured framework for creative exchange. Although the principle has traditionally guided human-human interaction, our results demonstrate that LLMs can effectively embody this approach, suggesting its potential as a broader framework for human-AI creative collaboration. Our controlled text-based environment builds on methodologies from joint action research (Noy, Dekel, and Alon 2011) and collaborative exploration studies (Rafner et al. 2023), providing a framework for systematic analysis of human-AI creative interaction.

Implications for AI System Design Our results suggest three key design principles for creative AI systems. First, LLMs' generative capabilities perform effectively in structured improvisational contexts, where creative invention is valued over factual precision. Second, interaction parameters, including timing, response length, tone, and willingness to explore human-initiated themes, need careful calibration to match natural human interaction patterns. Third, creative AI systems benefit from balancing narrative consistency with novel contributions. These insights point to AI systems designed not just for content generation, but for authentic creative partnerships.

Limitations Our study has four main limitations. First, the sample size (129 HH and 75 AIH stories) limits result generalizability. Second, our computational metrics relied on older models (GPT-2 and OPT-1.3B) and may miss important nuances of creative improvisation such as humor, emotional resonance, and advanced narrative techniques that expert evaluation might capture. Third, our participant pool consisted of non-expert improvisers, potentially affecting story creation and evaluation quality. Fourth, the text-only format eliminates important dimensions of face-to-face improvisation.

Conclusion

Key Findings Our study explores the potential of LLMs to participate in improvisational co-creation through the "Yes! and..." game. We found that stories co-created with AI partners received similar ratings to human-human collaborations in external evaluations, and participants in our "Turing test" experiment had difficulty distinguishing between the two conditions. These results suggest that LLMs can meaningfully engage in structured creative improvisation, contributing to coherent narrative development within the specific context of the "Yes! and..." game.

Methodological Contributions Our approach offers three key advantages for studying human-AI creative interaction:

- **Reduced Complexity:** The text-based format allows controlled model system for studying of core creative dynamics, similar to previous work in the domains of dyadic movement improvisation (Noy, Dekel, and Alon 2011) and visual exploration (Hart et al. 2017).
- **Quantifiable Assessment:** The experimental setup allows measurement of creative contributions through computational metrics.
- **Scalability:** Our chat-based platform enables systematic large-scale data collection and analysis.

Future Directions Our work points to several promising directions: comparing expert improvisers with AI performance, developing systems for dynamic AI contribution control in co-creative processes, implementing real-time evaluation systems for adaptive responses, and creating platforms for large-scale unbiased interaction studies. Beyond Turing test approaches, future work should explore AI's unique creative contributions rather than just human-like performance.

Broader Impact This work demonstrates the value of structured improvisational frameworks for human-AI creative collaboration. Our empirical evaluation shows that LLMs can participate in improvisational co-creation at levels comparable to human collaborators. By integrating improvisational techniques with computational creativity assessment, we establish a systematic framework for developing collaborative AI systems.

References

Abid, A.; Abdalla, A.; Abid, A.; Khan, D.; Alfozan, A.; and Zou, J. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild.

Agrawal, A.; Mackey, L. W.; and Kalai, A. 2023. Do language models know when they're hallucinating references? *ArXiv* abs/2305.18248.

Alon, U. 2010. Living within a story: Playback theatre and the art of improvisation. *Haaretz*.

Bates, J.; Loyall, A. B.; and Reilly, W. S. 1994. An architecture for action, emotion, and social behavior. In *Artificial Social Systems: 4th European Workshop on Modelling Autonomous Agents in a Multi-Agent World, MAAMAW'92 S. Martino al Cimino, Italy, July 29–31, 1992 Selected Papers 4*, 55–68. Springer.

Berliner, P. F. 2009. *Thinking in jazz: The infinite art of improvisation*. University of Chicago Press.

Boden, M. A. 2004. *The creative mind: Myths and mechanisms*. Routledge.

Branch, B.; Mirowski, P.; Mathewson, K.; Ppali, S.; and Covaci, A. 2024. Designing and evaluating dialogue llms for co-creative improvised theatre.

Branch, B.; Mirowski, P.; and Mathewson, K. W. 2021. Collaborative storytelling with human actors and ai narrators.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, 1877–1901. Curran Associates, Inc.

Bunescu, R. C., and Uduehi, O. O. 2022. Distribution-based measures of surprise for creative language: Experiments with humor and metaphor. In Ghosh, D.; Beigman Klebanov, B.; Muresan, S.; Feldman, A.; Poria, S.; and Chakrabarty, T., eds., *Proceedings of the 3rd Workshop on Figurative Language Processing (FLP)*, 68–78. Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics.

Chakraborty, T., and Masud, S. 2024. The promethean dilemma of ai at the intersection of hallucination and creativity. *Commun. ACM* 67(10):26–28.

Cho, H., and May, J. 2020. Grounding conversations with improvised dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2398–2413. Online: Association for Computational Linguistics.

Concepcion, E.; Gervas, P.; and Mendez, G. 2019. Evolving the ines story generation system: From single to multiple plot lines. In *ICCC*, 220–227.

Ely, J.; Frankel, A.; and Kamenica, E. 2015. Suspense and surprise. *Journal of Political Economy* 123(1):215–260.

Franceschelli, G., and Musolesi, M. 2024. Creativity and machine learning: A survey.

Gero, K. I.; Liu, V.; and Chilton, L. 2022. Sparks: Inspiration for science writing using language models. In *Proceedings of the 2022 ACM Designing Interactive Systems Conference*, 1002–1019.

Grace, K.; Maher, M.; Fisher, D.; and Brady, K. A. 2015. Data-intensive evaluation of design creativity using novelty, value, and surprise. *International Journal of Design Creativity and Innovation* 3:125 – 147.

Halperin, B. A., and Lukin, S. M. 2024. Artificial dreams: Surreal visual storytelling as inquiry into ai 'hallucination'. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*, DIS '24, 619–637. New York, NY, USA: Association for Computing Machinery.

Hart, Y.; Mayo, A. E.; Mayo, R.; Rozenkrantz, L.; Tendler, A.; Alon, U.; and Noy, L. 2017. Creative foraging: An experimental paradigm for studying exploration and discovery. *PLOS ONE* 12(8):1–15.

Hayes-Roth, B.; Sincoff, E.; Brownston, L.; Huard, R.; and Lent, B. 1994. Directed improvisation. *Knowledge Systems Laboratory Report KSL-94-61*.

Himberg, T.; Laroche, J.; Bigé, R.; Buchkowski, M.; and Bachrach, A. 2018. Coordinated interpersonal behaviour in collective dance improvisation: The aesthetics of kinaesthetic togetherness. *Behavioral Sciences* 8.

Hoffman, G., and Weinberg, G. 2010. Shimon: an interactive improvisational robotic marimba player. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, 3097–3102. New York, NY, USA: Association for Computing Machinery.

Jacob, M.; Chawla, P.; Douglas, L.; He, Z.; Lee, J.; Sawant, T.; and Magerko, B. 2019. Affordance-based generation of pretend object interaction variants for human-computer improvisational theater. In *ICCC*, 140–147.

Ji, Z.; Lee, N.; Frieske, R.; Yu, T.; Su, D.; Xu, Y.; Ishii, E.; Bang, Y. J.; Madotto, A.; and Fung, P. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.* 55(12).

Jiang, X.; Tian, Y.; Hua, F.; Xu, C.; Wang, Y.; and Guo, J. 2024. A survey on large language model hallucination via a creativity perspective.

Johnson, D.; Kaufman, J.; Baker, B.; Patterson, J.; Barbot, B.; Green, A.; van Hell, J.; Kennedy, E.; Sullivan, G.; Taylor, C.; Ward, T.; and Beaty, R. 2022. Divergent semantic integration (dsi): Extracting creativity from narratives with distributional semantic modeling. *Behavior research methods* 55.

Johnstone, K. 1987. *Impro: Improvisation and the Theatre*. London and New York: Routledge, 1st edition.

Jordanous, A. 2012. A standardised procedure for evaluating creative systems: Computational creativity evaluation based on what it is to be creative. *Cognitive Computation* 4.

Julian Just, Thomas Ströhle, J. F., and Hutter, K. 2024. Ai-based novelty detection in crowdsourced idea spaces. *Innovation* 26(3):359–386.

Karimi, P.; Grace, K.; Maher, M. L.; and Davis, N. 2018. Evaluating creativity in computational co-creative systems. *arXiv preprint arXiv:1807.09886*.

Koo, T. K., and Li, M. Y. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine* 15(2):155–163.

Kudrowitz, B., and Dippo, C. 2013. Getting to the novel ideas: Exploring the alternative uses test of divergent thinking. volume 5.

Lin, J. 1991. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory* 37(1):145–151.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. Roberta: A robustly optimized bert pretraining approach.

Liu, Y.-H.; Han, T.; Ma, S.; Zhang, J.; Yang, Y.; Tian, J.; He, H.; Li, A.; He, M.; Liu, Z.; Wu, Z.; Zhu, D.; Li, X.; Qiang, N.; Shen, D.; Liu, T.; and Ge, B. 2023. Summary of chatgpt/gpt-4 research and perspective towards the future of large language models. *ArXiv* abs/2304.01852.

Luchini, S.; Maliakkal, N. T.; DiStefano, P. V.; Patterson, J. D.; Beaty, R.; and Reiter-Palmon, R. 2023. Automatic scoring of creative problem-solving with large language models: A comparison of originality and quality ratings.

McKee, R. 1997. *Story: Substance, Structure, Style, and the Principles of Screenwriting*. New York: ReganBooks.

Nachmanovitch, S. 1991. *Free Play: Improvisation in Life and Art*. G.P. Putnam's Sons.

Noy, L.; Dekel, E.; and Alon, U. 2011. The mirror game as a paradigm for studying the dynamics of two people improvising motion together. *Proceedings of the National Academy of Sciences* 108(52):20947–20952.

Noy, L. 2014. *The Mirror Game: A Natural Science Study of Togetherness*. Bloomsbury Methuen Drama. 318–327.

Rafner, J.; Wang, Q. J.; Gadjacz, M.; Badts, T.; Baker, B.; Bergenholtz, C.; Biskjaer, M. M.; Bui, T.; Carugati, A.; de Cibeins, M.; Noy, L.; Rahimi, S.; Tylén, K.; Zana, B.; Beaty, R. E.; and Sherson, J. 2023. Towards game-based assessment of creative thinking. *Creativity Research Journal* 35(4):763–782.

Rezwana, J., and Maher, M. L. 2023. Designing creative ai partners with cofi: A framework for modeling interaction in human-ai co-creative systems. *ACM Transactions on Computer-Human Interaction* 30(5):1–28.

Ritchie, G. 2001. Assessing creativity. In *Proc. of AISB'01 Symposium*.

Shanahan, M. 2023. Talking about large language models.

Sowden, P. T.; Clements, L.; Redlich, C.; and Lewis, C. 2015. Improvisation facilitates divergent thinking and creativity: Realizing a benefit of primary school arts education. *Psychology of Aesthetics, Creativity, and the Arts* 9:128–138.

Swartjes, I., and Vromen, J. 2007. Emergent story generation: Lessons from improvisational theater. In *AAAI Fall Symposium: Intelligent Narrative Technologies*.

Theune, M.; Alofs, T.; Linssen, J.; and Swartjes, I. 2013. Having one's cake and eating it too: Coherence of children's emergent narratives. In *Workshop on Computational Models of Narrative*.

Thoppilan, R.; Freitas, D. D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; Li, Y.; Lee, H.; Zheng, H. S.; Ghafouri, A.; Menegali, M.; Huang, Y.; Krikun, M.; Lepikhin, D.; Qin, J.; Chen, D.; Xu, Y.; Chen, Z.; Roberts, A.; Bosma, M.; Zhao, V.; Zhou, Y.; Chang, C.-C.; Krivokon, I.; Rusch, W.; Pickett, M.; Srinivasan, P.; Man, L.; Meier-Hellstern, K.; Morris, M. R.; Doshi, T.; Santos, R. D.; Duke, T.; Soraker, J.; Zevenbergen, B.; Prabhakaran, V.; Diaz, M.; Hutchinson, B.; Olson, K.; Molina, A.; Hoffman-John, E.; Lee, J.; Aroyo, L.; Rajakumar, R.; Butryna, A.; Lamm, M.; Kuzmina, V.; Fenton, J.; Cohen, A.; Bernstein, R.; Kurzweil, R.; Aguera-Arcas, B.; Cui, C.; Croak, M.; Chi, E.; and Le, Q. 2022. Lamda: Language models for dialog applications.

Veale, T.; Wicke, P.; and Mildner, T. 2019. Duets ex machina: On the performative aspects of" double acts" in computational creativity. In *ICCC*, 57–64.

Yuan, A.; Coenen, A.; Reif, E.; and Ippolito, D. 2022. Wordcraft: Story writing with large language models. *27th International Conference on Intelligent User Interfaces*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; Du, Y.; Yang, C.; Chen, Y.; Chen, Z.; Jiang, J.; Ren, R.; Li, Y.; Tang, X.; Liu, Z.; Liu, P.; Nie, J.; and rong Wen, J. 2023. A survey of large language models. *ArXiv* abs/2303.18223.

Supplementary Information

Supplementary Information, including additional protocols, extended figures, and analyses, is available via a link at `https://tinyurl.com/yesand-si`.

Author Contributions

All authors participated equally in conducting the research and writing the manuscript.