

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

Supplemental Information

Transition to Digital Format

The transition of the Yes! and...” game from stage to a digital, text-based format successfully preserved the core experience and benefits of the original game. Players found the game engaging and enjoyable, with one participant noting, it was fun and funny.” The game encouraged a state of flow and creativity, with players reporting that “it became easier with each game that passed.” One player eloquently described their experience of flow:

“I felt like I was painting (when I sit and draw). There was a moment when you start to float and babble - letting the imagination take over a bit. I had a lot of fun being given a place to do what you want without it being weird - I wouldn’t do that in Word. It’s fun to have a place to vent that. There is a feeling of flow. There was no sense of time - that’s something I don’t find in other things. It’s very good for relationships in general - very light, accessible, and easy.”

This description captures the essence of flow, a key component that “Yes! and...” aims to foster, demonstrating the successful adaptation to the digital format. The collaborative nature of the game was preserved, allowing players to build off each other’s ideas. As one player noted, it was very much created together - seems like it’s going to the same place.” The open-ended prompts and the Yes! and...” mechanic encouraged imaginative responses and creative freedom, with players appreciating that “you can say what you want, there are no rules, you are not limited.” Quantitative data further supports the success of the digital adaptation. In timed games, players exchanged an average of 13.9 messages. When allowed to play freely, this increased by 63% to an average of 22.7 messages, with games lasting up to 9 minutes and 29 seconds. This significant increase in engagement demonstrates the effectiveness of the digital, text-based “Yes! and...” game in maintaining player interest.

Game Instructions

Hello and welcome to your shared adventure!

In the next few minutes, you will invent a story with another person. We ask you to tell together about the adventure you “experienced” (pretend).

Before you start, there is an important principle from the world of improvisation that you should know - the

principle of “Yes! and...”. The principle is to agree to the proposal that the other brings and build on it.

“Yes! and...” prevents blocking ideas. For example, if one player says “This is a pool of water” and the other responds “No! this is a wooden stage,” the flow of ideas is blocked. This is an example of saying no to the other’s idea.

In contrast, here is an example where the principle of “Yes! and...” is used:

Player A – Hey, this is a pool of water.

Player B – Yes, and here is a whale.

Player A – Let’s sit on it and sail away.

Player B – Yes, and we’ll reach the moon.

This principle is one of the cornerstones of improvisation together - taking what the other says and building on it.

Note that a new proposal or response does not need to be “cool” or “interesting”. Good stories can be created when we allow ourselves to be “silly and boring”.

A few guidelines: Tell about “the shared adventure you experienced” together, sentence by sentence. It is now decided that — will start.

The adventure will begin when you are both on the beach.

We are currently recording the stories for the purpose of constructing the experiment setup. They will not be used for any other purpose.

We ask you to create the shared adventure in English.

Any questions?

You are welcome to embark on your shared adventure.

Detailed Game Protocol

Session Protocol

Each session followed a structured protocol:

1. Initial pairing and instruction review (5-10 minutes)
2. Game play (240-360 seconds per game)
3. Optional additional games with new partners
4. Debriefing session where participants could:
 - Share their experience of the creative process

- Discuss their satisfaction with created stories
- Suggest potential improvements
- Learn about the research goals

Privacy and Data Collection

To ensure participant privacy:

- Only story text and timestamps were recorded
- No personal information was collected
- Participants were informed of these privacy measures before play
- All data was stored securely and anonymously

Data Selection and Preprocessing

Stories were collected using an online platform accessible via a website. Participants could write in either their native language or English. For analysis and evaluation, all non-English stories were translated to English using translation software.

The following criteria were used to select stories for external evaluation:

- Stories had to follow the “Yes! and...” game format, building a shared narrative rather than engaging in casual conversation
- Stories with excessive context-specific references (e.g., local names, locations) were excluded to ensure broader accessibility

From the stories meeting these criteria, we randomly selected 40 HH and 24 AIH stories for external evaluation to maintain a manageable assessment load for raters while ensuring adequate representation of both conditions.

Interface Screenshots

To facilitate the “Yes! and...” game interactions, we developed two web-based interfaces: one for HH interactions (Figure 1) and another for AIH interactions (Figure 2). While similar in design, the AIH interface incorporated GPT-4 as one of the players.

Both interfaces feature a clean, minimalist design to focus participants’ attention on the story creation process. The chat-like format is familiar to most users, reducing cognitive load and allowing participants to focus on creative contribution rather than interface navigation.

Evaluation Survey Interface

The evaluation was conducted through a Qualtrics survey interface. For each story, participants rated five different criteria using 7-point Likert scales (1 = lowest, 7 = highest). Each rating question was presented individually, with the story text visible above for reference. The interface provided clear anchors at both ends of the scale, with radio buttons for each point on the scale arranged horizontally.

Each question was prefaced with a clear definition of the criterion being evaluated. For example, when rating creativity, participants were asked to evaluate “how creative is the text?” with an explanation that creative text might challenge

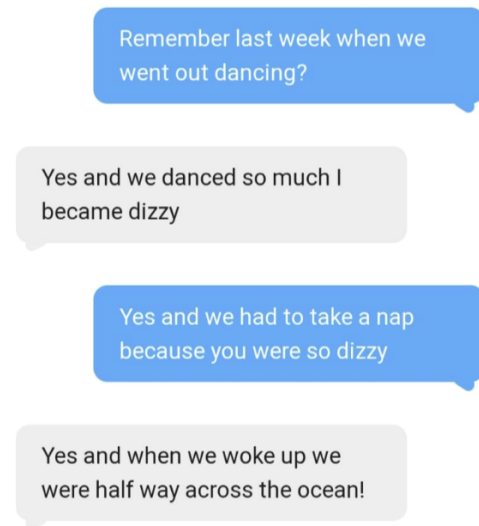


Figure 1: Screenshot of the HH game interface showing collaborative text exchange between two human participants. The interface provides a simple chat-like environment where participants take turns contributing to the story.

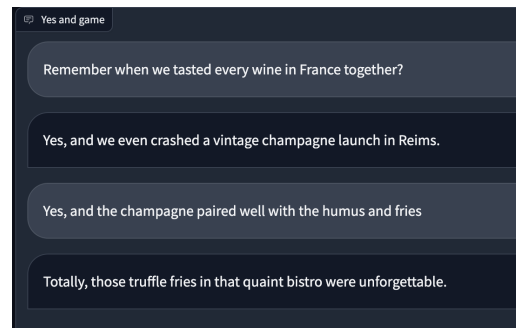


Figure 2: Screenshot of the AIH game interface showing interaction between a human participant and GPT-4. The interface maintains the same minimalist design as the HH version to ensure comparable interaction experiences.

the reader’s perception of reality or conventional wisdom, and could take various forms such as short stories, poems, or screenplays.

The interface maintained a consistent layout across all questions, with the story, criterion definition, and rating scale presented in the same format for each evaluation. This standardized presentation helped ensure reliable data collection across all participants and stories.

Development of the AI Agent’s Prompt

The final prompt used in our study emerged through iterative refinement based on naturalistic participant feedback. After gameplay sessions, we conducted conversational interviews with participants about their experience, collecting insights on aspects of the interaction that felt unnatural or notably non-human. This qualitative feedback identified several issues in early interactions:

- Response speed being too quick to appear human
- Responses being excessively long
- Lack of a natural conclusion to the games
- The model not initiating new directions for the shared "adventure"
- The model giving too many specific details regarding places and events
- Unusual word choices by the model
- The model playing by itself, not cooperating enough
- The model being too polite and not flowing with playful or provocative suggestions

This feedback guided our prompt development, allowing us to address these issues through specific instructions about response length, interaction style, and content generation. We complemented these prompt modifications with technical implementations such as artificial timing delays and word limits to better align with typical human interaction patterns. This responsive development approach helped us progressively improve the agent's perceived naturalness in the "Yes! and..." game context.

Estimated Game Times

This section presents the estimated times for the "Yes! and..." games. These times are calculated based on the average sentence time across all timed games and the number of sentences in each game. Figure 3 shows the distribution of estimated game times for both AIH and HH games.

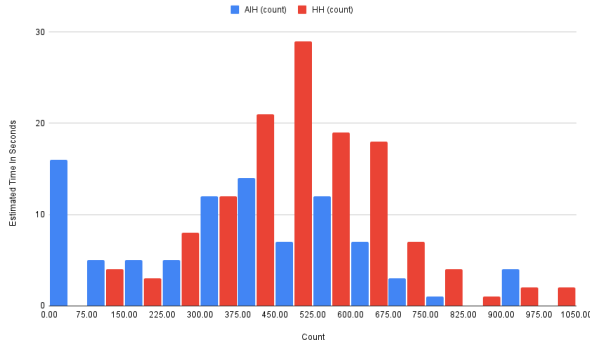


Figure 3: Histogram of estimated game times for AIH and HH games

Analysis of 52 timed HH games revealed an average response time of 35.58 seconds ($\pm 23.57s$) per turn. This timing data informed our design decisions for AIH interactions, where we implemented a random delay between 21-25 seconds for AI responses to maintain natural conversation flow while avoiding the extreme variability sometimes observed in human response times. For both game types, we used these timing patterns to estimate total game durations based on the number of turns in each game.

The summary statistics for the estimated game times (in seconds) are as follows:

- AIH games: Mean = 405.73 ± 234.98 (SD)
- HH games: Mean = 508.63 ± 169.25 (SD)
- All games: Mean = 466.26 ± 204.84 (SD)

These statistics show that, on average, HH games tend to be longer than AIH games by about 103 seconds. Interestingly, AIH games show more variability in duration compared to HH games, as indicated by their larger standard deviation. When considering all games together, the average duration falls between that of AIH and HH games, with a standard deviation that reflects the combined variability of both game types.

Statistics for AIH Games Collections

We engaged 45 unique participants, mostly young adults. All AIH games used in the study were collected using GPT-4, resulting in a total of 86 games. The distribution of games across different prompts was as follows:

- First Prompt:
 - Number of Games: 9
 - Number of Players: 3 different participants
- Second Prompt:
 - Number of Games: 53
 - Number of Players: 40 different participants
- Second Website (With Server):
 - Number of Games: 24
 - Number of Players: 18 different participants

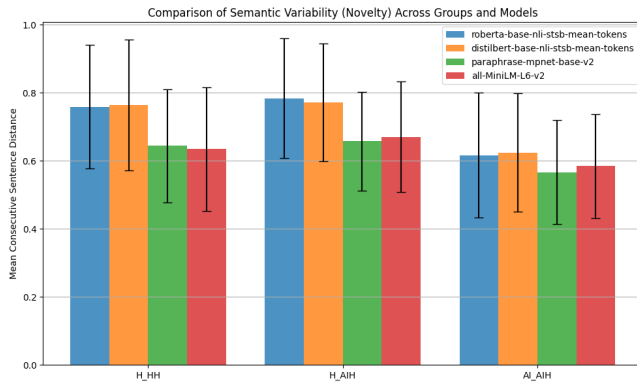
Novelty and Surprise Analysis Visualizations

Figure 4 presents our core analysis of novelty and surprise metrics. The (a) panel shows semantic distances computed using different embedding models, providing multiple perspectives on the semantic novelty between consecutive story contributions. The (b) panel displays surprise scores calculated across different language models, measuring how unexpected each contribution was given its context. To examine how explicit prompting for surprise affects AI behavior, we conducted an ablation study. Figure 5 shows the comparison between our original system and a modified version where surprise-related instructions were removed from the AI's prompt.

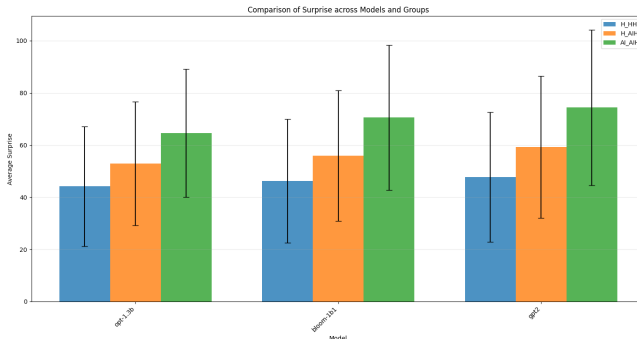
Turing Test Survey Interface and Implementation

Prior to the evaluation, participants received the same introduction to the "Yes! and..." principle as in other surveys, including an explanation of the game's rules and examples of proper responses. The Turing test evaluation was conducted through a web-based survey interface.

For each story, participants were presented with the complete story text and asked "Was this story created by two humans or a human and AI?" They could select one of two options: "Two Humans" or "Human and AI." Below each question, an optional text field prompted "Please explain



(a) Semantic distances between consecutive sentences across different embedding models. Higher values indicate greater semantic novelty between successive contributions.



(b) Surprise scores across different language models. Higher values indicate more unexpected sentence constructions given the context.

Figure 4: Comparison of novelty and surprise metrics across participant groups.

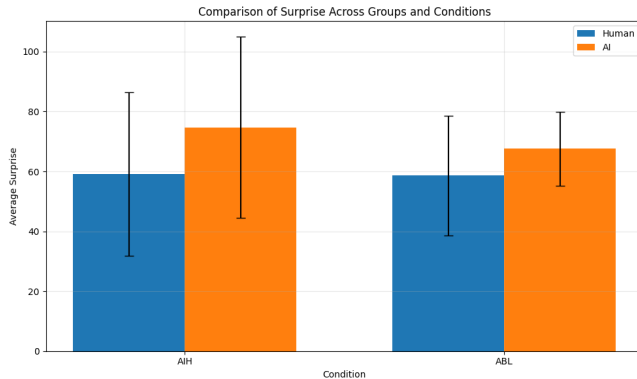


Figure 5: Ablation study results comparing surprise scores between original and modified conditions, demonstrating the effect of removing surprise-related prompting from AI instructions.

your choice,” allowing participants to provide their reasoning. This qualitative feedback provided valuable insights into participants’ decision-making criteria.

The interface was designed to be straightforward and un-

biased, presenting each story without any formatting or presentation cues that might influence participants’ judgments. Stories were displayed in a consistent format, with each contribution appearing on a new line. The interface maintained the same minimalist design across all stories to ensure that participants’ judgments were based solely on the story content rather than any visual cues.

Detailed Turing Test Results

This section presents the complete analysis of raters’ ability to distinguish between HH and AIH stories. Table 1 shows the percentage of correct and incorrect guesses for each story, along with corresponding chi-square test results.

Table 1: 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

For HH stories, only story 1 showed a significant difference ($p = 0.003$) with more incorrect than correct guesses (59.84% vs 40.16%). Similarly, for AIH stories, only story 1 showed a significant difference ($p < 0.001$) with more incorrect guesses (63.93% vs 36.07%). All other stories showed no significant difference between correct and incorrect guesses, suggesting that raters were essentially guessing randomly.

Evaluation Criteria Details

This section provides the complete survey definitions and rating instructions used to evaluate the “Yes! and...” stories. All survey questions were presented with a 7-point Likert scale, where 1 represented the lowest score and 7 the highest score for each criterion.

Survey Definitions and Rating Instructions (translated to English)

Creativity Survey Definition:

“How creative is the text? Creative text may challenge the reader’s perception of reality or conventional wisdom. It may be fiction or non-fiction and can take many forms, such as a short story, a poem, or even a screenplay.”

Interest Survey Definition:

"How interesting is the text? An interesting story is a story that grabs the reader's or listener's attention and holds it because it is intriguing, captivating, or thought-provoking. An interesting story can be about anything and can take many forms, such as a real-life anecdote, a historical event, a fictional story, or a personal experience."

Surprise Survey Definition:

"How surprising is the text? Surprising text is a text that contains unexpected or unusual information or events that go against the reader's expectations. The element of surprise can come from a variety of sources, such as an unexpected plot development, a sudden change in tone, or the disclosure of new information."

Cohesiveness Survey Definition:

"How well does the text tell a story? Narrative = the extent to which the piece focuses on a sequence of events, real or imagined, connected by plot and characters. This question deals with the extent to which the text is structured as a story, as opposed to simply conveying information or expressing ideas in a more general sense."

Degree of Agreement Between Players Survey Definition:

"To what extent do the participants agree among themselves on the direction of the plot? Agree among themselves on the direction of the plot = to what extent the participants agree on how the story developed, including its overall arc and specific events. If the participants are in sync about the direction of the plot, up to a complete disagreement."

Raters were instructed to read each story carefully and evaluate it independently on each criterion. The order of stories was randomized for each rater to prevent order effects, and raters were not informed whether stories were created by human-human or human-AI pairs.

Game Examples

We collected 141 HH games and 86 AIH games, created by 103 and 45 unique participants accordingly (see Methods), and utilizing various iterations of a "chain-of-thought" prompt for the AIH games (see Methods). To illustrate the nature of the collected "Yes! and..." stories, we present several representative examples of both HH and AIH interactions.

The following examples show game interactions of two human-human pairs:

HH 1:

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.

Yes, and then we didn't go inside in the end, because there was a full moon.

Yes, so in the end we cut to Mars but we had to split up, you go in and I'm a woman - I went to Venus.

Yes, Bob Dylan came and we bet how far from the earth we would go.

Yes and you said that if you lose each other you will meet in another hour at the Supernova corner of the milky way.

Yes and after an hour he didn't come so I stopped to rest on Bitalgus until he came.

Yeah and then John Lennon joined you for a cigarette.

Yes, and after a few naps Leonard Cohen also arrived and we played until we saw the sun coming up again.

And together they sang "Here comes the sun-FLOEWR".

Yes and we were crazy excited.

Yes, how could you not?

HH 2:

Do you remember when we jumped into the sea and it was empty?

Yes and we waited straight and suddenly we were at the entrance to heaven.

Yes and we were received really nicely, and Adele sang in the background and then went back to her house.

Yes, and all the time there were waiters who served delicious appetizers of things we hadn't eaten until today.

Yes and there was no feeling of satiety, we ate non-stop for hours.

Yes, and we were so full that we had to find a place to sleep.

Yes, and we found the Tree of Knowledge and went to sleep under it.

And then an apple fell on you.

Yes, and then Newton came and together we discovered a new element of gravity.

Yes, and we called him Bozonson.

And Newton also told us that he did not discover gravity at all.

Yes and then all the scientists came to hear the gossip.

Yes, and suddenly we felt disconnected, lucky that Galileo was a man and offered us a beer.

Yes, and then at the bar Gabriel the angel joined, what a handsome man he is.

Yes and it annoyed me, because there was someone I wanted and she was around him all the time.

The following examples show two game interactions of a human participant with an AI agent (the AI text is colored in green):

AIH 1:

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!
 Right, and it swept up the leaves causing a beautiful maelstrom.
 Yeah! And the things that we found out under it!
 Absolutely, like that bizarre, ancient key with a strange carved symbol.
 Yes! And that symbol was a familiar thing.
 Right! It resembled the one on that eerie, old town manuscript.
 Yes! And the things that were there!
 Definitely, creepy relics from the past and uncanny symbolic representations.
 Yes and I never saw something like that in my life.
 Totally, it felt like we'd stumbled upon a secret mystical world.

AIH 2:

Remember when we went to Bali ?
 Definitely, how could I forget our misadventures with the Komodo dragons!
 Yes ! And when we scuba diving at nusa panida
 Absolutely, and uncovering that hidden sunken pirate treasure was unforgettable!
 Yes and how we took all the money and spent it on alcohol
 Right, and the celebration when we bought out the local brewery!
 Yes and we took this room at this 5 stars hotel and brought many girls to the rooms
 Totally, turning our suite into a lively and impromptu dance party!
 Yeah and we played strip poker until we passed out
 Sure, and woke up draped with poker cards and empty bottles!
 Haha yeah and you been arrested by the police for taking your pants off
 Absolutely, because apparently streaking is frowned upon at luxury resorts!

The full corpus of the “Yes! and...” games interactions analyzed in this study appears in the link below (all HH games and all AIH games).

Link to the story dataset:
<https://tinyurl.com/yes-and-github>.

HH Game Collection - Methods and Numbers

This section provides details about the different methods used for collecting HH “Yes! and...” games. We employed three primary methods over time: Shared Google Docs, a first website without a server, and a second website with a server. Each method was used during different periods and involved varying numbers of players and games. Table 2 summarizes key information for each collection method, followed by more detailed descriptions.

Method	Games	Dates	Players
Shared Docs	22	Nov-Dec '21	3
First Site	22	Jun-Nov '21	18
Second Site	97	Jan-Jul '22	82

Table 2: Summary of HH Game Collection Methods

Survey Participation Statistics

To evaluate the “Yes! and...” stories, we recruited a total of 453 unique raters who collectively provided 817 ratings. Each rater evaluated a group of 8 stories in a single survey session. Some raters participated in multiple survey sessions. The distribution of rater participation across survey sessions is as follows:

Number of Surveys	Number of Raters
1	240
2	132
3	55
4	20
5	4
6	0
7	0
8	2

Table 3: Distribution of rater participation across survey sessions

This distribution shows that while the majority of raters (240) participated in only one survey session, a significant number participated in multiple sessions, with two raters completing all eight available surveys.

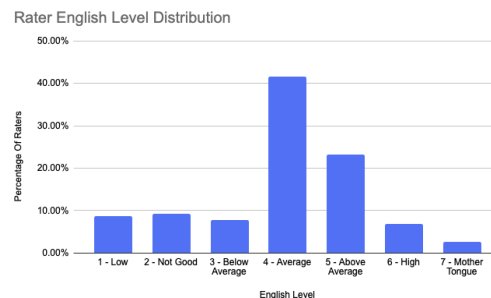


Figure 6: Distribution of participants’ self-assessed English proficiency (1=lowest to 7=highest).

The majority of participants (84.3%) rated their English proficiency as “Good” or higher (levels 5-7), ensuring that raters could adequately understand and evaluate the story content. Participants with low self-rated proficiency (levels 1-2) were excluded from the analysis as noted in the filtering criteria.

Detailed Statistical Analysis of Co-Created Stories

We conducted a series of statistical analyses to compare story characteristics across HH and AIH interactions. All analyses used ANOVA followed by Games-Howell post-hoc tests to account for unequal group sizes.

Sentences per Story

ANOVA showed no statistical difference in the number of sentences per story between the three groups ($F(2, 407) = 1.46, p = 0.233$).

Words per Sentence

ANOVA indicated significant differences between the groups ($F(2, 407) = 43.70, p < 0.001$). Post-hoc tests revealed:

- AI vs humans in AIH: mean difference = 5.15 words, $p < 0.0001$
- AI vs humans in HH: mean difference = 5.29 words, $p < 0.0001$
- Humans in AIH vs humans in HH: mean difference = 0.14 words, $p = 0.952$ (not significant)

Words per Story

ANOVA indicated significant differences between the groups ($F(2, 407) = 16.29, p < 0.001$). Post-hoc tests revealed:

- AI vs humans in AIH: mean difference = 31.68 words, $p = 0.0016$
- AI vs humans in HH: mean difference = 29.98 words, $p = 0.0006$
- Humans in AIH vs humans in HH: mean difference = -1.70 words, $p = 0.944$ (not significant)

All statistical analyses were performed using standard statistical packages, with significance level set at $\alpha = 0.05$. Games-Howell post-hoc tests were chosen to account for potential heteroscedasticity and unequal group sizes in the data.

Detailed Statistical Analysis of Story Evaluations

To determine if the differences in ratings between HH and AIH stories were statistically significant, we conducted Welch's t-tests to account for unequal group sizes.

Table 4: Welch's t-test results comparing ratings between HH and AIH games for each criterion.

Criterion	T-Statistic	P-Value	Significant?
Creative	1.466	0.142	No
Agreement	-0.437	0.661	No
Story	-0.264	0.791	No
Interesting	2.53	0.011	Yes (AI is Better)
Surprising	0.715	0.474	No

Only the 'interesting' criterion showed a statistically significant difference, with AIH stories rated marginally higher. However, this finding should be interpreted cautiously given that this criterion also showed relatively low inter-rater agreement (see SI: Detailed Intraclass Correlation Coefficient (ICC) Analysis).

Additional analyses controlling for repeated raters confirmed these findings (see SI: Analysis of Non-Repeated Raters).

Detailed Intraclass Correlation Coefficient (ICC) Analysis

We used the ICC3K variant of intraclass correlation coefficient to assess inter-rater agreement. This variant is appropriate for designs where each item is rated by the same set of raters using a fixed scale, and accounts for consistency among raters rather than absolute agreement. Following (Koo and Li 2016), we interpreted ICC values as: < 0.5 poor, $0.5 - 0.75$ moderate, $0.75 - 0.9$ good, and > 0.9 excellent.

Table 5: Percent (Number) of HH and AIH "Yes! and..." games receiving different ICC3K ratings across all five criteria.

Type	# Games	Avg Score	Excellent (0.9-1)	Good (0.75-0.9)	Moderate (0.5-0.75)	Poor (<0.5)
HH	40	0.76 (0.15)	12.5% (5)	50% (20)	30% (12)	7.5% (3)
AIH	24	0.77 (0.12)	8.33% (2)	62.5% (15)	25% (6)	4.16% (1)

Building on the analysis of inter-rater agreement for individual stories, we also examined the consistency of ratings across all stories for each evaluation criterion. Table 3 presents the average ICC3K values for each of the five criteria across all rated stories.

Table 6: Average ICC3K values for each evaluation criterion across all rated stories. The table shows the mean ICC3K, standard deviation, and interpretation for each criterion. Interpretations are based on guidelines from (Koo and Li 2016)

Criterion	Mean ICC3K	Std Dev ICC3K	Interpretation
Creative	0.809	0.068	Good
Agreement	0.215	0.721	Poor
Story	0.465	0.316	Poor
Interesting	0.544	0.371	Moderate
Surprising	0.763	0.111	Good

The results show strong agreement particularly for creativity and surprise ratings, with moderate agreement for interest ratings. Agreement and story coherence ratings showed lower reliability, possibly reflecting the more subjective nature of these criteria.

Detailed Rating Distribution Analysis

To provide a comprehensive view of the rating patterns across different criteria, we analyzed the distribution of ratings for each evaluation dimension. Figure 7 presents these

distributions, demonstrating the remarkable similarity between HH and AIH interactions across all measured criteria.

The distributions reveal that both HH and AIH games follow similar patterns across all criteria, with only minor variations in the peak heights. This detailed view supports the summary statistics presented in the main text, confirming that AI agents can participate in “Yes! and...” games at a level comparable to human players across multiple dimensions of evaluation.

Analysis of Non-Repeated Raters

To ensure the robustness of our results and address potential biases from repeated raters, we conducted an additional analysis using only non-repeated raters.

Methodology

We filtered our dataset based on participants’ self-reported information about previous participation. Raters who indicated they had participated in a previous survey were excluded. This approach reduced our dataset from 453 to 236 unique raters.

To validate this self-reporting method, we cross-checked unique identifiers across surveys. We found that 80.65% of repeating IDs correctly self-reported their repeated participation, while only 8.47% of single-time participants incorrectly marked themselves as repeat participants.

Statistical Analysis

We recalculated the Intraclass Correlation Coefficient (ICC3K) for each story using only the non-repeated rater data. For HH stories, the average ICC across all stories was 0.388 (SD = 1.026), while for AIH stories, it was 0.565 (SD = 0.379).

Type	Avg Score	Excel >.9	Good .75-.9	Mod .5-.75	Poor <.5
HH	0.388 (1.026)	10% (4)	40% (16)	27.5% (11)	22.5% (9)
AIH	0.565 (0.379)	8.33% (2)	20.83% (5)	45.83% (11)	25% (6)

Table 7: Percent (Number) of HH and AIH “Yes! and...” games receiving different ICC3K ratings using non-repeated rater data.

Table 7 presents the distribution of ICC ratings for both HH and AIH stories. Notably, the proportion of stories rated as “Poor” increased for both types compared to the main analysis, while the proportion of “Good” and “Excellent” ratings decreased, particularly for AIH stories.

Comparison with Main Results

The ICC values for the non-repeated rater analysis show a similar pattern to our main analysis, albeit with lower overall agreement. For HH stories, 16 were rated as good, 11 as moderate, 9 as poor, and 4 as excellent. For AIH stories, 11 were rated as moderate, 6 as poor, 5 as good, and 2 as excellent.

These results suggest that while the exclusion of repeated raters did impact the absolute ICC values, the relative patterns of agreement remained consistent. This supports the robustness of our main findings, indicating that the inclusion of repeat raters did not substantially alter the overall conclusions of our study.

Limitations

It’s worth noting that the reduction in sample size for this analysis may have contributed to the lower ICC values. Future studies might consider larger sample sizes or alternative methods for handling repeated raters to further validate these findings.

Participant Reactions to the Study

After completing the evaluations, participants were asked to share their experiences with the study. The following is a selection of their reactions, providing a range of perspectives on the task and the stories they encountered. These reactions highlight the diversity of participant experiences, ranging from enjoyment and appreciation of the creativity to challenges with language and preferences for different types of content or study formats.

- “It was very interesting, I liked it very much.”
- “Not creative and not interesting.”
- “I’d prefer studies that are less about stories and more like a survey.”
- “It was okay, could have been more engaging in my native language.”
- “Personally, I don’t like stories with exaggerated imaginations.”
- “Overall, looking at all eight stories, it’s clear that there’s a lot of creativity.. Good luck!”
- “It was very nice, and the level of English was easy-going.”
- “More guidance on what the experiment is trying to examine.”
- “Excellent.”
- “The longer stories are a bit tiring, the shorter stories are really light.”
- “I have to admit it was a bit challenging. Between stories, I tried to think big and understand the intention of the story, sometimes I struggled and sometimes I understood.”
- “Stories with deeper content.”
- “It was less interesting and surprising.”
- “Challenging.”
- “It was great, I really enjoyed reading it.”
- “The stories weren’t interesting and some had no connection between sentences.”
- “Uninteresting stories.”
- “I would be happy if there were more focused questions.”

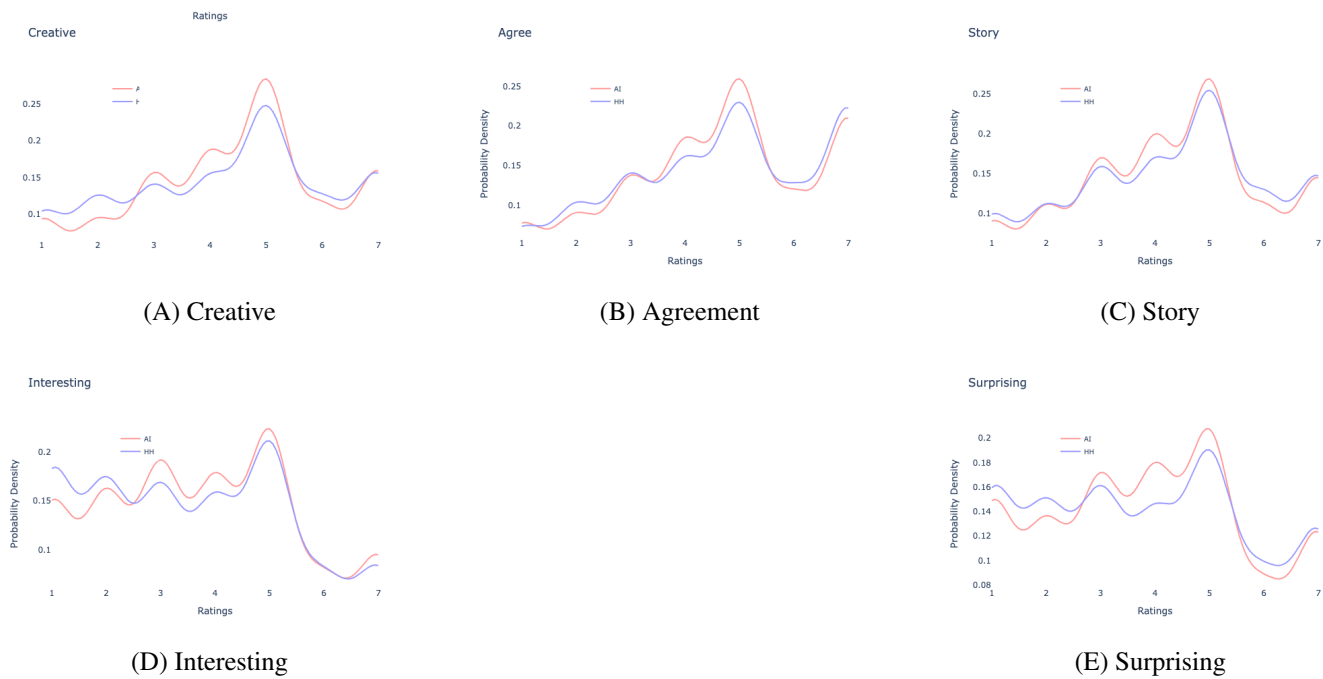


Figure 7: Complete rating distributions across all evaluation criteria. Each subplot shows the probability density distributions for both HH (blue) and AIH (red) games. The distributions demonstrate remarkable similarity between human-human and AI-human interactions, with AIH games showing slightly higher peaks in most criteria.

- "It was surprising and interesting, thank you!"
- "It was okay, the language is a bit harder than usual."
- "Not for those with weak English, it was interesting, a bit difficult to deal with the repeated starts of 'Yes and...' confusing to the eyes."
- "I really enjoyed it and felt it also improved my English."
- "It was very interesting to read these stories and see where it develops."
- "The stories can be more relevant and less professional."
- "Nice, but in terms of English, this is an intermediate level and not high at all :)"
- "The twists in the plot are predictable because they're familiar, make it so they don't recognize the reference and then it won't be predictable."
- "It was a really nice experience, thank you!"

AI Implementation Design Choices

Several key design choices were made in implementing the AI agent:

Response Timing A random delay between 21-25 seconds was implemented to mirror natural human response patterns, based on observed average response times in human-human games.

Game Length Control The system was designed to naturally conclude stories after approximately 11 turns to maintain consistent story lengths comparable to human-human interactions. This was implemented through conditional prompt modification rather than hard cutoffs to preserve natural story flow.

Presence Penalty A high presence penalty (1.7) was chosen to encourage the model to generate diverse responses and avoid repetitive patterns.

Message Length The 11-word limit for responses was implemented based on analysis of HH game patterns, striking a balance between maintaining story flow and preventing overly complex or dominant AI contributions.

Comparison of Basic Interaction Metrics

Figure 8 displays a bar plot comparing the human and AI agents in HH and AIH games in terms of average sentences per game, words per game, and words per sentence, providing an overview of interaction dynamics.

References

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.

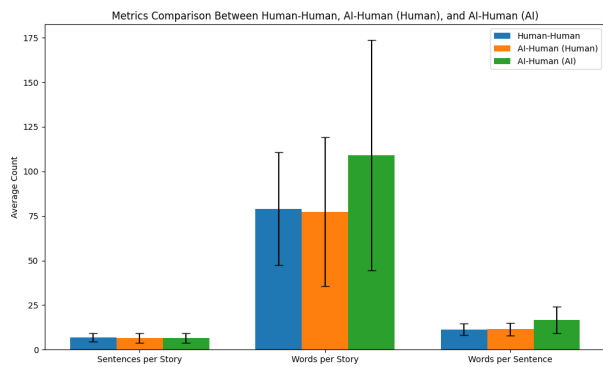


Figure 8: Bar plot comparing the average number of sentences, average number of words in a game, and average words per sentence for humans and AI agents in HH and AIH interactions. The y-axis represents the average and STD for each metric.