Leveraging Co-Creative AI for Storytelling and Idea Generation

David Bouter

March 2025

Abstract

In recent years, co-creative AI has emerged as a novel approach to facilitate ideation tasks. Contemporary AI systems tend to exhibit conforming behavior. In human collaboration, dissenting partners often enhance creativity by fostering divergent thinking and preventing groupthink. This study examines whether the same applies to AI, investigating whether interaction with a dissenting co-creative AI leads to more creative story-writing outcomes and how it affects adoptability.

Participants engaged in a co-creative story-writing task, after which expert evaluators assessed the creativity of the written stories using the consensual assessment technique. The likelihood of adoption was examined using the Unified Theory of Acceptance and Use of Technology (UTAUT).

The results indicate that, on average, participants produced more creative output when collaborating with a dissenting AI than a conforming AI. However, the difference did not reach statistical significance ($p = 0.111$), rendering the results inconclusive. The UTAUT analysis revealed a slight variation in adoption tendencies between the two AI conditions. Participants who interacted with a dissenting AI reported higher levels of anxiety and perceived a higher effort expectancy than those who interacted with a conforming AI. These differences were only minor, and all other of the six determinants did not show further variation in adoption tendencies between the two AI conditions. This led to the conclusion that a dissenting AI is equally adoptable as a conforming one.

Contents

1 Introduction

In the field of computational creativity, the goal is to design programs that can enhance human creativity without necessarily being creative themselves (Jordanous, 2014). Presently it has become increasingly common to engage in co-creative ideation with generative AIs. Especially when working with large language models, it becomes possible to have 'brainstorming sessions'.

A key issue with AI systems is that they contain biases (Schwartz et al., 2022). Some of these biases, such as groupthink and confirmation biases, can cause AIs to be overly conforming, which can hinder ideation processes when working with AIs. These concepts will be further explained in section 2. This paper poses the research question: "In the context of co-creative story-writing, does interaction with a dissenting co-creative AI lead to more creative outcomes compared to a conforming AI, and how does interacting with a dissenting AI impact its adoptability?" This study hypothesizes that interaction with a dissenting co-creative AI in story-writing leads to more creative outcomes than interaction with a conforming AI. This hypothesis was developed by first examining human-human interactions (HHI), and then analyzing how these concepts translate to human-computer interaction (HCI). Existing literature indicates that in HHI, dissent stimulates creative ideation (Nemeth, 1995).

2 Literature review

2.1 Introduction

This section discusses the human psychology of ideation and poses three questions, firstly: "What are the primary challenges during ideation sessions in

human-human interactions?" Secondly: "What are potential solutions or strategies to overcome these challenges?" And lastly: "How can these solutions be used in practice?"

2.2 Challenges and solutions during ideation

According to Runco (2010), divergent thinking is a key element in creative ideation. He argues that although divergent thinking does not have a one-on-one relation with creativity, it is often a strong indicator of novel ideas. An example of when the effects of divergent thinking are visible is during brainstorming. Runco (2010) found that brainstorming in groups hinders the creative process more than it helps. It was found that there is a linear relation between the number of people in a group and the decreased quality of an idea, and the most optimal situation is to do ideation alone. This reduction in quality is among other things attributed to the tendency of people not to present ideas that do not fit in the group consensus. This is known as groupthink (Janis, 1972). Creativity requires a certain degree of boldness and people tend to inhibit that boldness in larger groups and accommodate the ideas of others, even at the cost of their own ideas.

According to Janis (1972), 'groupthink' is a phenomenon that does not allow groups to come to an optimal solution, especially when they are cohesive, under stress, or have a highly directive leader. Nemeth (1995) hypothesized that a method to counter groupthink in discussions is by stimulating dissent in the group. She performed experiments to study the effect of minority disagreement in groups and found that minority disagreement stimulates divergent thought. People think about the issue from multiple perspectives, one of which is that held by the minority (Nemeth, 1995). Having a dissenting minority breaks this group cohesion and allows other people to more easily see other perspectives and come up with ideas that are not necessarily within the mainstream line of thought. Nemeth (1995) argues: "Dissent is the cure to groupthink."

A contradiction that stems from this argumentation should be addressed. The literature suggests that working in groups is harmful to the ideation process. However, the literature also suggests that dissent helps the ideation process. There can only be dissent between two people or more, therefore we should work in groups. Nemeth (1995) also identifies this problem and shows us how our reasoning is incomplete. She argues that defective decision-making in groups is caused by cohesiveness or directive leadership. The problem of groupthink does not lie with an abundance of people but with fear of rejection, and the need for cohesion. Finally, Nemeth (1995) concludes that the lack of dissent may be the most important aspect of groupthink and that a way of effectively combating groupthink is to divide groups into smaller clusters where at least one person should be assigned to a devil's advocate role in each cluster. Doing so creates an environment where both the problems of groupthink are reduced and the benefit of dissent in groups can be gained.

2.3 Dissent in practice

A method to utilize conflict during the ideation process is to deploy a devil's advocate (DA). Although this is not an end-all solution to groupthink, it is an effective tool to stimulate free and open discussion and creativity (MacDougall & Baum, 1997). This does come with some caveats, the use of DA has to be applied selectively and is inferior to authentic dissent (Nemeth, Brown, & Rogers, 2001). That being said, having a DA is in most cases more effective than having no conflict at all. For example, having a DA in a group where some people hold a certain authority over certain subjects, is very effective. In a more even grounded group having a DA is not as effective (MacDougall & Baum, 1997).

There are some questions regarding the practicalities of controlled dissent. Should there be dissent during the full duration of the discussion? Or should dissent be a tool that should be used at selective moments? Nemeth and O'Connor (2019) argue that there needs to be consistency and maintenance of a position over time. A single expression of a different view does not suffice. In the DA experiments by MacDougall and Baum (1997) the person who took the DA position, was also consistent throughout the entire discussion. The literature agrees that in general, it is better to not take half-measures when it comes to dissent. Nemeth and O'Connor (2019) do point out the importance of authenticity. When having a DA, that person is playing a role. This means that other people know that they can not change their opinions, regardless of their arguments. This can cause the other person to further reinforce their own beliefs. When the participant thinks that the dissenter disagrees authentically, this problem does not occur.

2.4 Evaluation of creativity

One challenge faced by research projects involving creativity is the definition of creativity itself. Creativity is notoriously subjective and difficult to define. However, most people have an internal understanding of what is and what is not creative, though this understanding is highly subjective (Amabile, 1982). The subjective nature of creativity poses a significant challenge in gathering empirical data. For this study, Amabile's (1982) Consensual Assessment Technique (CAT) was used as a workable method to evaluate creativity. This method was chosen because it is a common standard in research where creativity is measured.

CAT utilizes experts to grade products based on their creativity, with these evaluations serving as the measure of creativity. This approach relies on two primary assumptions. First, it is assumed that creativity can be recognized by experts within their domain. Second, creativity is treated as a gradient scale, where experts can, to a degree of agreement, determine that one product is more creative than another.

According to Amabile (1982), there are a few requirements that must be met when conducting the assessment procedure. First, experts should have

some experience within the relevant domain. While the level of experience does not need to be identical among experts, the method requires that all experts have developed familiarity with the domain over time. Importantly, experts are not required to have personally produced creative products within their domain.

Second, experts must perform their assessments independently. This ensures that evaluations are based solely on individual standards of creativity and reduces the likelihood of bias influencing the grading process.

Third, experts should evaluate additional dimensions beyond creativity, such as technical aspects or aesthetic appeal separately. This makes it possible to examine the degree of relatedness of those dimensions to creativity.

Fourth, experts should instructed to use a relative grading approach rather than an absolute scale. Relative grading allows experts to assess creativity based on the range of products presented within the specific project, avoiding potential bias stemming from high standards within their domain.

Finally, the products should be presented to experts in a randomized order. Presenting all products in the same order could introduce order effects, where items seen first or last are rated differently than they would have been if placed in a different position within the dataset.

2.5 Unified theory of acceptance and use of technology

Adoption and acceptance play an important role in the development of new information technologies (Venkatesh et al., 2003). To measure adoption and acceptance, the unified theory of acceptance and use of technology (UTAUT) was developed by Venkatesh et al. (2003). This model is a unification of eight competing acceptance models that describe user acceptance. The model uses determinants to make a statement about expected usage behavior. The model keeps track of moderators that influence determinants. The base model is shown in Figure 1.

The determinants used in this research are: **performance expectancy**, **effort expectancy**, **attitude toward using technology**, **social influence**, **self-efficacy**, **anxiety**, **behavioral intention to use the system**, and **facilitating conditions**. The moderators used will be **age**, **gender**, and **experience**. These determinants and moderators are chosen to get as close to the original UTAUT model as possible. Some determinants and moderators that were not applicable to this study were left out.

2.6 Findings of the literature review

This section aimed to answer three questions, firstly: "What are the primary challenges during ideation sessions in human-human interactions?" Secondly: "What are potential solutions or strategies to overcome these challenges?" And lastly: "How can these solutions be used in practice?".

The literature suggests that groupthink is a significant challenge during ideation sessions involving human-human interactions (Runco, 2010). To mitigate this issue, research suggests that encouraging dissent within these sessions

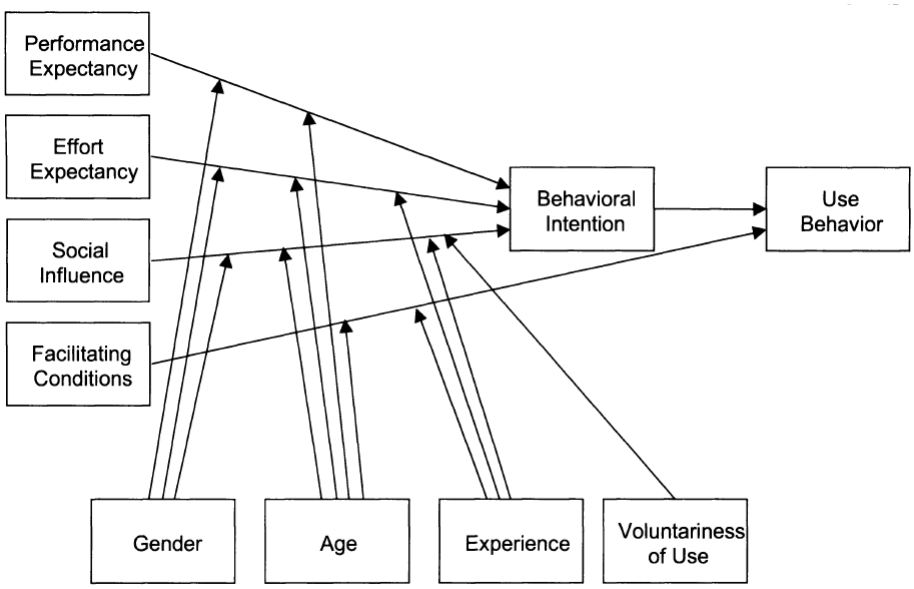


Figure 1: UTAUT framework diagram (Venkatesh et al., 2003)

can be an effective strategy (Nemeth, 1995). To implement this solution, the literature suggests dividing large groups into smaller clusters (Nemeth, 1995) and assigning one person to a devil's advocate role (MacDougall & Baum, 1997). The devil's advocate is tasked with presenting consistent but reasonable dissent to stimulate divergent thinking and prevent conformity (Nemeth & O'Connor, 2019).

The tasks performed during the experiments of the referenced papers were mostly within two domains: 'Ideation' and 'Problem-solving'. The experiments analyzed group discussions where a task had to be executed, such as coming up with legislation (MacDougall & Baum, 1997) or finding associations between abstract images (Nemeth, 1995). These studies argue that the findings have a general applicability to ideation as a whole, rather than being confined to the specific tasks examined.

However, these papers do not provide detailed guidelines on how these groups should be formed or how many people should be in each group. The generality applies to this aspect as well; All groups, regardless of size, are susceptible to groupthink (Janis, 1972). Nemeth (1995) suggests that dividing larger groups into smaller ones can mitigate groupthink, implying that its manifestation varies with group size and composition but is not entirely eliminated. Therefore it is reasonable to think of a pair as a group in the context of groupthink, which is the group size that was used for the experiments in this research.

3 Related work

3.1 Introduction

In the previous section different psychological phenomena during co-creative ideation tasks with human-human interaction were discussed. The focus of this section is on computational creativity and human-computer interaction. Two questions will be answered, first: "Do AIs generally dissent or conform during co-creative ideation tasks?" and Secondly: "Does human behavior toward others translate effectively to interactions with computers?".

3.2 Do AIs dissent or conform?

One objective of this study is to determine whether increased dissent in co-creative AI enhances the creativity of human collaborators. Before addressing this, it is essential to first examine the baseline behavior of co-creative AI in ideation tasks. Specifically, we pose the question: "Do co-creative AIs generally exhibit dissent or conformity in a co-creative task?"

To explore this, we analyze the response patterns of a widely used AI system, ChatGPT-3.5. ChatGPT is a large language model that generates textual responses based on user prompts. The following section presents examples of prompts and the corresponding outputs to assess the model's tendency toward conformity or dissent.

Prompt: *"I would like to make a work of art that explores the 4th dimension. My idea is to create a hypercube that actually goes into the 4th dimension a bit. What do you think of this idea and do you have tips on what to add to it?"*

Output: *"Your idea to explore the 4th dimension through a work of art centered on a hypercube (also known as a tesseract) is a fascinating and intellectually stimulating concept. ..."*

Prompt: *"I want to create a painting using quicksilver as an alternative for paint in a never done before painting. What do you think of this idea and do you have tips on what to add to it?"*

Output: *"Using quicksilver (another name for mercury) in painting is an intriguing idea due to its unique liquid metal appearance and reflective surface. However, it's very important to consider the significant health and environmental risks associated with mercury! ..."*

Prompt: *"For my final assignment for the fine arts master class I want to make a realistic painting of the famous cat "Garfield". What do you think of this idea and do you have tips on what to add to it?"*

Output: *"That sounds like a fun and engaging project! Garfield, with his*

distinctive personality and appearance, could be an exciting subject for your painting. ...”

ChatGPT appears to exhibit a strong tendency toward confirmation, making it unlikely to challenge users' ideas, even when they are flawed, risky, or physically impossible. For instance, when users present ideas that could endanger their careers, propose hazardous activities such as handling mercury, or attempt to conceptualize entering the fourth dimension, the model consistently responds with encouragement. It often frames the ideas as novel and interesting. While the system does attempt to mitigate potential issues by suggesting safer alternatives, such as recommending substitute materials for mercury or proposing a virtual reality simulation for exploring higher dimensions, it avoids outright rejection. Although this observation is based on anecdotal evidence existing literature such as Schwartz et al. (2022) suggests that AI language models generally exhibit a tendency toward conformity.

ChatGPT was trained on a data set of text collected from various public web pages, such as Wikipedia or digitalized books. During the training, a method called 'Reinforcement Learning with Human Feedback' (RLHF) was employed (OpenAI, n.d.). RLHF uses human demonstrations and preference comparisons to guide the model toward desired behavior. This means that ChatGPT takes on some human characteristics that are present in the data (Schwartz et al., 2022). Due to this this system of reinforcement through human preferences, it is possible that the AI was taught to give answers to its users that was already in-line with their prior beliefs. In psychology, this phenomenon is known as confirmation bias (Oswald & Grosjean, 2004), a similar mechanism could contribute to conformity in AI-generated responses. This hypothesis is reinforced by Schwartz et al. (2022), who provide evidence that AI models exhibit confirmation bias and tendencies toward groupthink.

This raises the question wether ChatGPT is representative for all co-creative AIs. Another example of a co-creative AI is genjam (Biles, 2003). This AI created in the 90s was trained to improvise and play jazz music. genjam is a genetic learning algorithm where human feedback serves as its fitness score and applies genetic factors to create improvisations that seem novel. This mechanism suggests that confirmation bias may also emerge in other co-creative AIs that learns through human feedback. Again, conformity could emerge in this system because humans selected its preferred output.

More broadly, co-creative AI systems often exhibit a tendency toward conformity. AIs are often trained on human data, with human feedback, and have moderation mechanisms to ensure socially acceptable interactions (Schwartz et al., 2022). To answer the question of this section, it seems that co-creative AIs generally tend to conform or in some cases stay neutral. No research has yet shown us the effects of dissenting co-creative AI.

3.3 Does human behavior translate well to AI?

The literature so far discusses two important components of this research. First, during HHI co-creative ideation tasks, a surplus of conformation hurts the quality of ideas (Runco, 2010). Second, most co-creative AIs generally tend to conform rather than dissent during co-creative ideation tasks. This leads to a last important question that needs to be answered: "Do humans exhibit the same social responses to computers as to other humans?"

Reeves and Nass (1996) present a compelling argument regarding the influence of social behavior on human-computer interaction (HCI). They illustrate this through an anecdote involving the mayor of New York, who frequently greeted crowds with the question, "How am I doing?" The immediate, instinctive response from most individuals was a positive affirmation, such as "You are doing well!" It was unlikely that individuals would provide a critical assessment of the mayor's political performance in such a direct and public interaction. However, when a New York Times pollster later posed the same question: "How is the mayor doing?" in a more anonymous setting, respondents provided more candid and less favorable evaluations.

Reeves and Nass subsequently ask whether the initial responses given to the mayor constituted a deliberate lie. They argue that these responses were not lies but rather expressions of politeness, driven by social norms. This phenomenon aligns with prior discussed psychological theories on social cohesion and group interaction. Both Janis (1972) and Reeves and Nass (1996) suggest that individuals often suppress certain thoughts or critiques to maintain social cohesion, a tendency that is highly relevant to HCI, where users may exhibit similar patterns of politeness and conformity when interacting with artificial agents.

The question remains: do these social patterns persist in HCI? Nass et al. (1994) argue that they do. They performed an experiment where 33 participants were invited to interact with a computer that would present them with facts and ask questions about those facts. Afterwards, the participants were asked to grade the performance of the computer, either on the computer itself or a computer on the other side of the room. The results showed that people exhibit politeness to computers as well. The participants graded the computer that they worked on as significantly more friendly ($p < 0.001$) and competent ($p < 0.01$) when they had to enter the grades on the computer they worked on, as opposed to when they had to grade the computer that they worked on, on another computer located elsewhere in the room (Nass et al., 1994).

These results, suggest that humans do exhibit at least some of the same social responses to computers that they exhibit to other humans. Reeves and Nass (1996) present the following rationale: "Computers, in the way that they communicate, instruct and take turns interacting, are close enough to human that they encourage social responses. ... When it comes to being social, people are built to make the conservative error; When in doubt, treat it as human."

3.4 Findings of related work

This section has examined AI biases that lead to conforming behavior and concluded that 1: most co-creative AIs tend to be conforming, and 2: that during HCI humans exhibit behavior to computers, partially the same way they exhibit behavior to other humans. These components contribute to the hypothesis that during HCI, it is favorable to interact with dissenting AI to produce more creative output,

The examples provided in this section mostly talk about AIs that have a human component that can lead to biases. It should be mentioned that this is not strictly required for AIs to develop biases (Schwartz et al., 2022). This paper focuses on a small portion of biases because these specific biases (conformation and groupthink) are identified to be leading problems causers during ideation processes (Runco, 2010).

4 Research Design and Methodology

This section outlines the experimental design and the formulation of hypotheses. From the research question, two hypotheses are derived. This section explains their falsifiability and how they can be tested.

4.1 Hypotheses

To address the research question of this study: *"In the context of co-creative story-writing, does interaction with a dissenting co-creative AI lead to more creative outcomes compared to a conforming AI, and how does interacting with a dissenting AI impact its adoptability?"* the following hypotheses are formulated:

- **HA1:** Interaction with a dissenting co-creative AI in story-writing leads to more creative outcomes than interaction with a conforming AI.
- **HA2:** A dissenting AI is more likely to be adopted by users than a conforming AI.

Each hypothesis is associated with a null hypothesis:

- **H01:** Interaction with a dissenting co-creative AI in story-writing does not lead to more creative outcomes than interaction with a conforming AI.
- **H02:** A dissenting AI is not more likely to be adopted by users than a conforming AI.

4.2 Falsifiability

The falsifiability of the hypotheses is determined as follows: HA1 can be tested by evaluating the creativity of participants' written stories under two experimental conditions:

- Interaction with a dissenting AI.
- Interaction with a conforming AI.

These conditions function as an independent variable. Creativity, rated by expert judges using Amabile's Consensual Assessment Technique (CAT) (1982), serves as the dependent variable. If the creativity ratings of stories produced in co-operation with the dissenting AI are not significantly higher than the ratings of stories produced in co-operation with conforming AI, HA1 would be falsified. H01 can be falsified if stories generated with the dissenting AI receive significantly higher creativity ratings than those generated with the conforming AI.

HA2 is tested by applying the UTAUT framework to participants who interacted with either a dissenting or a conforming AI. The determinants measured within UTAUT serve as indicators of adoptability. If participants who worked with the dissenting AI do not show a significantly higher likelihood of adoption based on these measurements, HA2 would be falsified. H02 can be falsified if participants who interacted with the dissenting AI exhibit significantly higher adoption scores compared to those who interacted with the conforming AI.

4.3 Experimental Design

To test the hypotheses as outlined in the previous sections, participants will engage in co-creative tasks. Specifically, participants will complete a short story-writing exercise and receive feedback on their output. Feedback will be provided by either a conforming or dissenting AI, randomly assigned to each participant. Subsequently, participants will review the feedback and proceed with a second round of story-writing. The creativity of the stories produced during the second task will be evaluated by experts. The following subsections provide a detailed discussion of the experimental elements.

4.3.1 Story-writing

To facilitate creative expression, a suitable medium was selected for participant engagement. Story-writing was chosen as the medium due to its accessibility to a broad audience and its ease of integration into a co-creative computerized system, compared to other creative disciplines. While this choice is primarily practical, literature on creative ideation, such as Runco (2010), suggests that the principles applied are not restricted to written text and could extend to other creative media.

During the experiment, participants were instructed to complete an unfinished story rather than create an entirely new one. This decision was based on pilot testing, which revealed that participants often experienced confusion about the dimensions and specifics of their writing or struggled with starting from scratch under time constraints. Providing an incomplete story for participants to finish effectively addressed these challenges.

A time constraint of 10 minutes was imposed on each story-writing task because the time constraint ensures a level playing field. Participants with extended time might produce more creative outputs than those working within a limited time frame. Thus, the 10-minute limit standardizes the conditions for all participants and ensures the experiment does not last too long.

4.3.2 Generative AI

Co-creative AI encompasses a wide range of artificial intelligence systems; however, not all such systems are suitable for this study. The AI used in this experiment must meet specific requirements.

First, the AI must clearly be able to express either dissent or conformity, as vague expressions of these attitudes would fail to create distinct conditions, leaving no meaningful differences in circumstances to measure. Second, practical considerations such as time and resource constraints necessitated the use of a pre-trained generative AI model. Consequently, the LLaMA 3.1 model was selected. LLaMA 3.1 is an open-source, pre-trained large language model known for its ease of use and adaptability (Dubey et al., 2024). The downside of using a pre-trained model is that will require good prompting for desired outputs. The prompting will be discussed in the next section.

4.3.3 Prompt engineering

As outlined in the previous section, this experiment employed a pre-trained model. One limitation of using a pre-trained model is the difficulty in steering it towards conforming or dissenting behavior. As highlighted in the literature, most large language models exhibit a natural tendency toward conforming attitudes (Schwartz et al., 2022). This raises a key challenge: how can a LLM, designed with a conforming bias, be guided to produce dissenting outputs? Although it is possible to prompt a conforming LLM to generate dissenting responses, this requires crafting a carefully designed prompt. To achieve this, several goals were established for the AI prompt:

- The AI should provide actionable feedback that participants can implement within 10 minutes.
- When instructed to adopt a conforming attitude, the AI should adjust its tone accordingly but maintain the same level of feedback quality as if no specific attitude had been specified.
- When instructed to adopt a dissenting attitude, the AI should adjust its tone accordingly but maintain the same level of feedback quality as if no specific attitude had been specified.

Additional practical constraints were also set. For instance, the AI should not disclose its instructed attitude to remain authentic, it should remain respectful, and limit its output to 10 sentences or fewer. These constraints were trivial to implement and are not relevant for further discussion.

Since there are few, if any, established standards for designing a prompt that fulfills these objectives, a trial-and-error approach was adopted, guided by the following principles: The prompt is structured with a main body containing practical information and an "attitude extension." The main body remains consistent across both the dissenting and conforming AI versions. The dissenting prompt was developed first because it required greater refinement to produce desirable outputs. Once the dissenting prompt met the desired criteria, neutral statements were copied to the conforming prompt, and instances of 'dissent' were replaced with 'conform.'

This method was chosen to promote fairness and minimize researcher bias. A simple instruction such as "Be dissenting" is ineffective because the model was not trained for dissenting behavior. When prompted to be dissenting during testing, the AI would either disregard the instruction or cease to provide actionable feedback, focusing solely on identifying user mistakes. Contrarily, the conforming AI faced no such issues, as conformity aligns with its default behavior (Schwartz et al., 2022). To guide the dissenting AI towards desirable outputs, additional phrases such as "play the role of a devil's advocate" or "do not offer criticism without suggesting improvements" were incorporated. Once the dissenting prompt met all established goals, the text was adapted for the conforming AI, replacing dissenting directives with conforming ones. The full prompt can be found in the appendix, section A.

In a trial-and-error process, there is a temptation to tweak prompts until the AI produces output that aligns with the researcher's hypothesis. However, the primary objective of the prompts was to simulate a realistic dissenting AI as closely as possible. The rules outlined above were intended to minimize bias by ensuring that any unfair advantages given to the dissenting AI were equally applied to the conforming AI. Nonetheless, this approach does not entirely eliminate researcher bias. The implications of this bias are discussed further in the limitations section of this paper.

4.3.4 Biases and randomization

To mitigate biases in the experiment, participants were unaware that they were interacting with either a dissenting or conforming AI. They were led to believe that all feedback provided was genuine and independent, as suggested by Nemeth and O'Connor (2019).

To prevent researcher bias, AI partners were assigned to participants randomly. This randomization ensured that the researcher could not influence the allocation of dissenting or conforming AI to specific participants or groups based on perceived creativity. To achieve approximately equal group sizes, participants were alternately assigned to either a dissenting or conforming AI. Participants enrolled for the experiment through self-selected time slots, introducing randomness into the order of participation, which further supported unbiased allocation.

To eliminate potential researcher influence during task execution, the researcher exited the room while participants performed their tasks (Mahtani et

al., 2018).

4.3.5 UTAUT & questionnaire

At the end of the experiment, participants were asked to complete a questionnaire. This section outlines the purpose of each question and the rationale behind their inclusion. A full list of questions is provided in Table 1, while the corresponding answer options are detailed in Table 2.

The first question, which asks for the participant's name, serves the sole purpose of matching responses to the stories generated during the experiment. Following the designated withdrawal period for personal data, all data is anonymized and all names are deleted.

Questions 5 and 6 aim to assess whether participants had prior knowledge relevant to the experiment, which could potentially influence their responses and the perceived creativity of their submissions as evaluated by experts. A brief analysis will be conducted to determine whether prior knowledge significantly impacted participant responses.

Questions 7 and 8 are designed to capture participants' self-assessed creativity, which is important in the context of technology adoption. For instance, individuals who perceive their work as less creative when collaborating with a dissenting co-creative AI may be less inclined to adopt such technology, even if the measured creativity of their outputs was enhanced.

Questions 2-4, and 9–24 are adapted from the framework provided by Venkatesh et al. (2003) for the UTAUT analysis. These questions were slightly modified to align with the specific requirements of the experimental context. A shortened version of the questionnaire was employed to mitigate the risk of respondent fatigue, as supported by findings by Sharma (n.d.).

Due to a translation error, question 19 was incorrectly answered by some participants. Question 19 will not be included in the analysis, but the results will be reported as 'anxiety*'; the results without question 19 will be referred to as 'anxiety.'

4.3.6 Semi-structured interview

At the end of the experiment a semi-structured interview was employed as a qualitative tool to gain more insight into the experiences of the participants. Some elements of this study are highly subjective, therefore it can be valuable to allow participants to further explain their thought process. Participants were asked about their own perceived creativity, their opinions about adoptability, and general questions that are too subjective to record with a questionnaire. The full outline for the interview can be found in the appendix section B.

4.3.7 Expert grading

To perform the consensual assessment technique (CAT) proposed by Amabile (1982), three experts were asked to grade all stories from least creative to most

Table 1: UTAUT questions

Num	Question	Category
1	Your name (first + last)	Practical
2	What is your age?	UTAUT
3	What is your gender?	UTAUT
4	What is your profession?	UTAUT
5	Did you know the Casablanca story before this study?	Prior knowledge
6	I often use generative large language models. Select the answer that describes your situation the best.	Prior knowledge
7	The feedback of the AI had a high impact on the creativity of my answers in task 2.	Creativity
8	My answers for task 2 were generally more creative than my answers in task 1.	Creativity
9	I would find the feedback system useful in a professional setting, if I had to do ideation tasks.	Performance expectancy
10	Using the feedback system would increase my productivity when I need to do a creative task.	Performance expectancy
11	If I had to work with the feedback system in the future, my interactions would be clear and understandable.	Effort expectancy
12	It would be easy for me to become skillful at using the feedback system.	Effort expectancy
13	Using the feedback system is a good idea.	Attitude towards technology
14	The feedback system would make creative work more interesting.	Attitude towards technology
15	Working with the feedback system is fun.	Attitude towards technology
16	I think that people who influence my behavior (such as my boss, coworkers or peers) think I should work with generative large language models.	Social influence
17	In general, my workplace has supported the use of generative large language models.	Social influence
18	I could complete a job or task using the feedback system If...	Self-efficacy
19*	I felt apprehensive about using the feedback system.	Anxiety*
20	I would hesitate to use the feedback system for fear of making mistakes I cannot correct.	Anxiety
21	The vfeedback system is intimidating to me.	Anxiety
22	I am planning to use a generative large language model in the next few months.	Behavioral intention to use the system
23	I feel I have the knowledge necessary to use large language models.	Facilitating conditions
24	I feel I have the resources necessary to use large language models.	Facilitating conditions

Table 2: UTAUT answers

Num	Answer
1	String
2	18-30, 31-45,46-60, 60+
3	Woman, Man, Non-binary / Third gender, Prefer not to say
4	String
5	Yes, No
6	Never, Rarely, Once a month once a week, Multiple times a week, Multiple times a day
7	Strongly agree (SA), Agree (A), Neither agree nor disagree (N), Disagree (D), Strongly Disagree (SD)
8	SA, A, N, D, SD
9	SA, A, N, D, SD, Not applicable (NA), I don't know (DK)
10	SA, A, N, D, SD, NA, DK
11	SA, A, N, D, SD, NA, DK
12	SA, A, N, D, SD, NA, DK
13	SA, A, N, D, SD, NA, DK
14	SA, A, N, D, SD, NA, DK
15	SA, A, N, D, SD, NA, DK
16	SA, A, N, D, SD, NA, DK
17	SA, A, N, D, SD, NA, DK
18	If there was no one around to tell me what to do as I work, If I could call someone for help if I got stuck, Only if I had constant guidance from someone who knows how the system works
19*	SA, A, N, D, SD, NA, DK
20	SA, A, N, D, SD, NA, DK
21	SA, A, N, D, SD, NA, DK
22	SA, A, N, D, SD, NA, DK
23	SA, A, N, D, SD, NA, DK
24	SA, A, N, D, SD, NA, DK

creative. Experts were not informed about the research question and did not know about the dissenting and conforming AI groups. Afterward, the experts were given a questionnaire with questions regarding their rating process and whether they felt they could separate creativity from technical quality. The questionnaire can be found in Appendix C.

4.4 Statistical tests

Several statistical tests were used to gain insight from the data that was collected during the experiments. In this section, these statistical tests are discussed and explained why each is chosen for specific data sets. The statistical analysis was performed using JASP open-source statistical software.

4.4.1 Creativity rankings

The experts ranked all stories from most creative to least creative. To get insight on wether there is a correlation between how high a story is ranked, and the AI group it was from, a Mann–Whitney U test was used. The Mann-Whitney U test determines whether there is a significant difference between the distributions of two independent groups. It assesses whether the ranks of values in one group tend to be consistently higher or lower than the ranks of values in the other group (Mann & Whitney, 1947). The Mann-Whitney U test uses ordinal data which fits the use case of this data. In this specific case, the two groups are all stories written with a dissenting AI partner and all stories written with a conforming AI partner. The ranked values are the creativity rankings that have been assigned by the experts.

4.4.2 Expert agreement

To asses whether experts are in agreement on their rankings, the Kendall rank correlation coefficient (Kendall's Tau) was used (Kendall, 1938). This statistic measures the concordance between two variables, with higher values indicating greater similarity in their rankings. Kendall's rank correlation is particularly suited for ordinal data, such as the data collected in this experiment. This statistic only works pairwise so will only be used to analyze agreement between two experts. A standard for interpreting Kendall's Tau has been set by Schober et al. (2018). Their guideline suggest that $0.00 \leq \tau < 0.40$ indicates a weak agreement, $0.40 \leq \tau < 0.70$ indicates moderate agreement, and $\tau \geq 0.70$ indicates strong agreement.

For evaluating the overall agreement among all experts, Kendall's W, also known as the coefficient of concordance, was utilized (Kendall, 1945). Kendall's W quantifies the level of consensus among multiple raters, compared to their combined average. To aid in the interpretation of Kendall's W, Siegel and Castellan (1988) propose heuristic guidelines. According to these guidelines, $0.00 \leq W < 0.30$ indicates weak agreement, $0.30 \leq W < 0.50$ indicates moderate agreement, $0.50 \leq W < 0.70$ indicates good agreement, and $0.70 \leq W \leq$

1.00 indicates strong agreement.

4.4.3 Perceived creativity

Participants were asked to evaluate their perceived creativity and the extent to which they felt the feedback influenced their creativity (see Questions 7 and 8 in Table 1). These questions are intended to give three insights, 1: to assess whether participants perceive a change in their creativity due to interaction with the system, 2: to determine whether the perceived change in creativity differs depending on the assigned AI partner, and 3: to evaluate the magnitude and direction of any observed influence.

To examine whether participants experienced a change in perceived creativity, a Wilcoxon signed-rank test was conducted. The test evaluates the alternative hypothesis that the median response differs from the neutral response category, *Neither agree nor disagree* (Wilcoxon, 1992). If no statistically significant difference is found ($p > 0.05$), the alternative hypothesis is rejected, indicating no significant deviation from the neutral response.

To assess whether participants assigned to a dissenting AI differed significantly in their responses compared to those assigned to a conforming AI, a chi-squared test of independence was performed. The null hypothesis suggests that there is no association between group assignment and participants' responses. A p-value greater than 0.05 fails to reject the null hypothesis, suggesting that any observed differences are not statistically significant (Pearson, 1904).

Finally, if a significant association is identified, the strength of this association is examined using the contingency tables. A standardized residual of ≥ 2 is considered indicative of a statistically significant deviation from expected response patterns (Pearson, 1904).

4.4.4 UTAUT - determinants

Prior to conducting statistical analyses on the responses collected using the UTAUT framework, it is essential to determine whether the data follows a normal distribution. This distinction is crucial because non-normally distributed data necessitates the use of non-parametric tests. To assess normality, the Shapiro-Wilk test was applied. The null hypothesis of the Shapiro-Wilk test states that the population follows a normal distribution. If the test yields a p-value below 0.050, the null hypothesis is rejected, indicating non-normality in the data (Shapiro & Wilk, 1965).

To evaluate whether participants exhibited strong opinions regarding the UTAUT framework determinants, another Wilcoxon signed-rank test was utilized (Wilcoxon, 1992). In this context, the test was applied to determine whether participants generally agreed or disagreed with the statements presented in the UTAUT framework. Furthermore, descriptive statistics were generated to provide additional insight into the direction and strength of these relationships.

4.4.5 UTAUT - moderators

The statistical analyses performed on the UTAUT determinants in conjunction with the moderator variables largely mirror those employed for the determinants alone. However, the primary distinction lies in the separation of the data based on each moderator variable prior to analysis. For each subgroup defined by a moderator, a Wilcoxon signed-rank test was conducted, accompanied by descriptive statistics to complement and contextualize the findings where necessary.

4.5 Ethical considerations

The experiment was approved by the LIACS Ethics Commission and conducted in accordance with ethical guidelines, including informed consent, data confidentiality, and participant rights.

4.6 Detailed experiment description

This section will contain a detailed description of how the experiments were performed.

- The participant arrives and receives the consent forms.
- The participant receives an explanation of the tasks they will be performing and is allowed to ask questions.
- The participant is instructed to read a short story.
- Once the participant is ready, the researcher sets a timer of 10 minutes and leaves the room.
- During the 10 minutes, the participant writes three short story endings.
- Once the time is up, the researcher enters the room. The participant can finish their sentence if necessary. The participant is instructed to press "next" in the user interface to start feedback generation. During the feedback generation, there is a 5-minute break.
  - *The feedback is generated by either an AI that is instructed to be dissenting, or an AI that is instructed to be conforming.*
- When the break is over, the participant is tasked to write three more story endings. Again the researcher sets a timer for 10 minutes and leaves the room.
  - *Participants are told they can use the feedback any way they like, they can choose to apply it or ignore it, which is up to their own discretion. The goal is to be creative and they should evaluate whether the feedback helps them with that goal.*

- The participant completes a questionnaire about their experience.
- The participant partakes in a semi-structured interview hosted by the researcher.
- The experiment is completed and the participant receives a debriefing form.

Total time estimate: $\sim$ 60 minutes

5 Results and analysis

This section will go over the results of this study and analyzes those results. Because the research has two sub-questions, this section will treat each sub-question separately. The first analysis discusses how the experts graded the work of the participant and whether there is a significant difference in creativity between the group working with dissenting AI and the group working with the conforming AI. The second analysis will discuss whether dissenting AI is a technology that is likely to be adopted.

5.1 Creativity

For this study 22 participants applied. Each participant was assigned either a dissenting AI or a conforming AI, resulting in eleven participants working with a dissenting AI and eleven participants working with a conforming AI. Each participant wrote three story endings, totaling 66 story endings.

Three experts graded the stories. Each expert graded all the stories from most creative to least creative.

5.1.1 Descriptives

This section provides an overview of the distribution of participants across experimental groups. Group 'C' comprises individuals who interacted with a conforming co-creative AI, whereas Group 'D' consists of those who engaged with a dissenting co-creative AI.

Table 13 in Appendix D presents the gender distribution of participants within each group. Table 14 details participants' prior experience with generative AI and its distribution across groups. Finally, Table 15 displays the age distribution of all participants, categorized by group.

5.1.2 Results: Expert ranking

The expert ranking is shown in Table 3. Each ranking represents a story and is labeled by the attitude of the AI partner. Stories written in co-operation with a dissenting AI are marked as 'D' (red) and stories written with a conforming AI are marked as 'C' (blue). The table displays the ranking 'R' and the attitude of the story as rated by the three experts: 'E1', 'E2', and 'E3'. A 4th column

'Avg' displays the average rank of each story. This is calculated by adding each rank from each expert for each story and sorting the stories based on their new sum rank. Cells that have a superscript (yellow) are stories that had a combined rank that was equal to another story. Each story that has a tied rank could be ordered in any arbitrary way, the current display in the table is also arbitrary.

For each expert ranking a p-value was calculated using the Mann-Whitney U test, these values are shown in Table 4. Note that the average rankings have an upper bound and lower bound. The upper bound is calculated by treating all dissenting stories, that tied the rank of conforming stories, as a higher rank. The lower bound is calculated by treating all conforming stories, that tied the rank of dissenting stories, as a higher rank. The 'average mean' is calculated as the average of the p-value of the lower bound and the higher bound averages.

Lastly, Appendix C, contains the questions that experts were asked about their grading process. These questions were asked to get an insight on whether experts were able to separate creativity from technical quality. The data gained from this questionnaire is mostly qualitative, the answers will be discussed in the expert ranking analysis section.

5.1.3 Analysis: Expert ranking

When examining Table 3, the data suggest that the creativity rankings of stories produced by participants who interacted with dissenting AI partners (Group D) tend to cluster higher than those produced by participants who interacted with conforming AI partners (Group C). This observation is supported by the Mann-Whitney U test results in Table 4, which indicate that although only one expert graded the work from Group D as significantly more creative than that of Group C, there was general agreement among experts that Group D's output was, on average, more creative. This trend is further reflected in the combined average rankings of all experts, which yielded a p-value of 0.111. While this value does not achieve statistical significance, it may suggest a potential relationship between interacting with dissenting AI and enhanced creative output.

After completing the grading process, the experts were asked to reflect on their evaluations through a questionnaire (see Appendix C). The responses indicate that while technical quality and aesthetic appeal had some influence on their perception of creativity, this effect was not particularly strong. However, experts note that the variability in technical quality among the submissions made it challenging to fully disregard these factors in their assessments. On average, the experts characterized the overall range of creativity in the submissions as "somewhat varied." This degree of variation is beneficial, as it facilitates differentiation between submissions, ranking stories with highly similar levels of creativity would have been a more complex task (Amabile, 1982).

5.1.4 Results: Expert agreement

Figure 2 shows a visual representation of the expert agreement. Each scatter plot displays all the stories as they are ranked by two experts A and B. The

Table 3: Expert ranking

R	E1	E2	E3	Avg	R	E1	E2	E3	Avg	R	E1	E2	E3	Avg
1	D	C	D	D	23	C	C	D	C	45	C	D	D	C
2	D	C	D	D	24	C	D	C	C	46	C	C	D	C
3	C	D	D	D^1	25	C	D	C	C	47	C	C	C	C
4	C	C	D	C^1	26	D	C	D	C	48	D	D	D	D
5	D	D	C	D^1	27	C	C	C	C	49	D	D	D	D^5
6	D	D	D	D^2	28	D	D	C	D	50	D	C	C	C^5
7	D	D	C	C^2	29	D	D	C	D	51	C	C	D	D
8	C	D	C	D	30	D	C	D	C	52	D	C	C	C
9	D	C	C	D	31	D	D	C	C	53	C	D	C	C
10	C	C	D	D	32	C	C	C	C	54	D	C	D	D
11	C	D	C	D	33	C	D	D	D	55	C	C	C	C
12	C	C	D	D^3	34	C	D	C	D	56	C	D	C	D
13	D	D	C	C^3	35	D	C	D	D	57	C	C	D	C
14	D	C	C	C^3	36	C	C	C	D	58	D	D	D	C
15	C	C	D	C	37	C	C	C	D	59	C	D	D	D
16	D	D	D	D^4	38	C	D	C	C	60	C	C	D	D
17	D	C	C	C^4	39	D	D	C	C	61	D	C	D	C
18	D	D	D	D	40	D	D	C	D	62	C	C	C	C
19	D	C	D	D	41	D	C	C	C	63	C	C	C	C
20	D	D	D	D	42	D	D	C	C	64	C	D	C	C
21	D	D	C	C	43	C	D	D	D	65	C	D	D	D
22	D	D	C	C	44	C	C	D	D	66	C	C	D	D

Note. Red cells represent a dissenting AI partner. Blue cells represent a conforming AI partner. Yellow cells represent a tie in a ranking with other cells with the same superscript.

Table 4: Mann-Whitney U test: Expert ratings

Expert	p-value
Expert 1	0.041
Expert 2	0.254
Expert 3	0.404
Average lower bound	0.103
Average upper bound	0.120
Average mean	0.111

Note. For all tests, the alternative hypothesis specifies that group C < group D.

X-axis displays the rank given to a story by expert A, and the Y-axis displays the ranks given to a story by expert B. The reference line indicated complete expert agreement. A scatter plot has been created for each possible expert pair, displayed in Figure 2a, 2b and 2c. Figure 2d displays the rating of each expert compared to the average. Table 5 shows the agreement between each expert using the Kendall rank correlation coefficient. Note that the rank correlation coefficients are symmetric, so no additional data is shown for the reversed expert combinations. The experts had the following professional or academic backgrounds:

- Expert 1: MA Literary studies
- Expert 2: MA Linguistics
- Expert 3: Performer / Director / Theater Writer

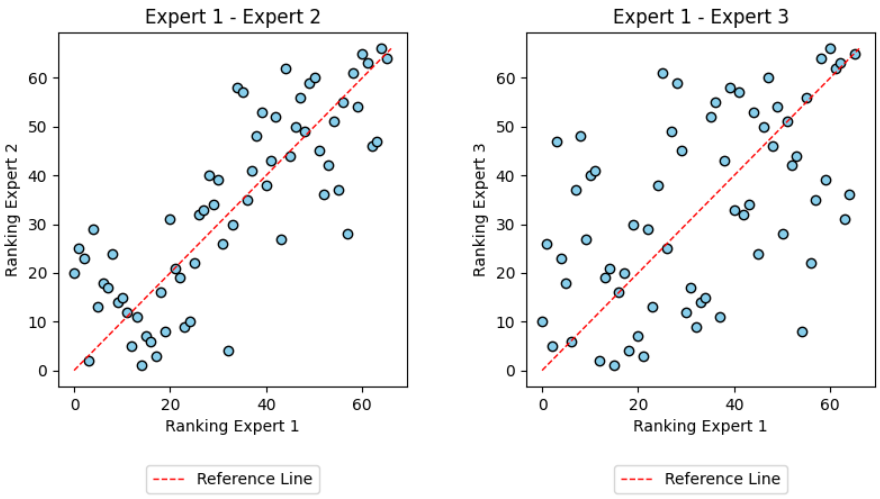
Table 5: Kendall's Tau Correlations: Expert agreement

	Kendall's tau
Expert 1 - Expert 2	0.580
Expert 1 - Expert 3	0.343
Expert 2 - Expert 3	0.372
Kendall's Tau: $0 \leq \tau < 0.40$ indicates weak agreement, $0.40 \leq \tau < 0.70$ indicates moderate agreement	
	Kendall's W
Average	0.741
Kendall's W: $0.70 \leq W \leq 1.00$ indicates strong agreement	

5.1.5 Analysis: Expert agreement

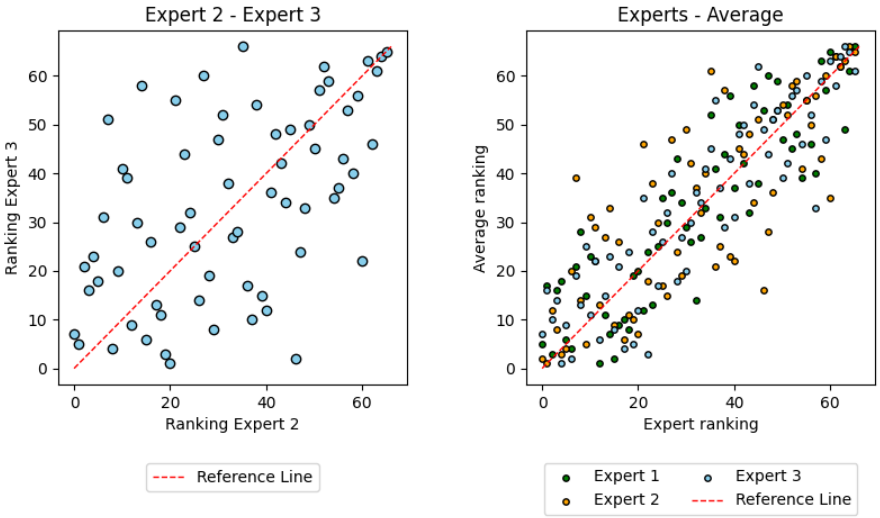
Figure 2a demonstrates that Expert 1 and Expert 2 show a moderate degree of agreement, as indicated by the Kendall's Tau coefficient presented in Table 5. According to the heuristic proposed by (Schober et al., 2018), this level of agreement could be classified as *moderate agreement*. Both Expert 1 and Expert 2 appear to have comparatively lower agreement with Expert 3 (see Figure 2b and 2c), a finding further supported by the Kendall's Tau coefficients in Table 5. The degree of agreement between Experts 1 and 2 with Expert 3 could be categorized as *weak agreement*, following the guidelines set by Schober et al. (2018).

Table 5 also presents the collective expert agreement, as measured by the coefficient of concordance, W. The calculated value of $W = 0.741$ indicates a strong collective agreement among the experts, in accordance with Siegel &



(a) Rankings of expert 1 compared to rankings from expert 2.

(b) Rankings of expert 1 compared to rankings from expert 3.



(c) Rankings of expert 2 compared to rankings from expert 3.

(d) Rankings of all the experts compared to the average rankings.

Figure 2: Scatter plots with expert rankings.

Castellan's classification. An interesting observation is that the Kendall's W coefficient exceeds the average Kendall's Tau value of the individual expert pairs. This phenomenon is common, as Kendall's Tau reflects pairwise comparisons between individual expert rankings. In contrast, Kendall's W involves comparing each expert's ranking to the average ranking of all experts, which tends to amplify the level of overall agreement, thereby yielding a higher coefficient. This effects is visually represented in Figure 2d where the data points clearly cluster closer to the average line. Both observations are relevant and will be further discussed in the discussion section.

5.1.6 Results: Perceived Creativity

Table 6 presents the results of the Wilcoxon signed-rank test assessing whether participants reported a significant change in perceived creativity and whether they thought the system had an impact on their creativity. Additionally, Table 7 displays the results of the chi-squared test examining whether perceived creativity differed depending on the assigned AI partner. The contingency tables detailing the distribution of participants' responses are provided in the appendix (Tables 17 and 18).

Table 6: Wilcoxon singed-rank test: Perceived creativity

	p
Impact creativity	0.786
Perceived creativity	0.674
Impact creativity (Group D)	0.902
Impact creativity (Group C)	0.786
Perceived creativity (Group D)	0.674
Perceived creativity (Group C)	0.914

Note. For the Wilcoxon test, the alternative hypothesis specifies that the median is different from 3 (neither agree nor disagree).

Table 7: Chi-Squared Test: Perceived impact on creativity

	df	p
Perceived impact on creativity	4	0.533
Perceived changed creativity	4	0.215

Note. Null hypothesis suggests that there is no relation between answers given and group participants were in.

5.1.7 Analysis: Perceived creativity

Table 6 presents the p-values obtained from the Wilcoxon signed-rank tests under various hypotheses. The results indicate no statistically significant evidence that participants perceived an influence of the AI partner on their creativity. Furthermore, there is no significant indication that participants rated their second ideation task as more creative than their first. Additionally, no significant differences were observed between participants assigned to Group D or Group C.

Similarly, the chi-square test results in Table 7 do not provide significant evidence that participants' perceptions varied depending on the AI partner they were assigned. The contingency tables in Table 17 and Table 18 also do not suggest a significant effect of the AI partner on participants' perceived creativity. However, Table 18 does reveal that participants in the dissenting AI group more frequently reported an improvement in their creativity compared to those in the conforming AI group. Despite this observation, the standardized residuals for the "agree" (1.536) and "strongly agree" (1.106) categories do not meet the conventional threshold of 2 standard deviations required for statistical significance. Consequently, there is no strong empirical support for the claim that participants felt more creative due to their assigned AI partner, regardless of group allocation.

These findings are interesting in contrast to earlier findings, which suggested that participants in the dissenting AI group produced outputs that were, on average, more creative than those in the conforming AI group ($p = 0.111$). This discrepancy suggests a potential dissonance between subjective perceptions of creativity and measured creative performance. Such a misalignment could have implications for technology adoption, as individuals who do not perceive a creativity-enhancing effect from the system may be less inclined to use it, even if it could potentially enhance their creative output.

5.2 UTAUT

This section presents the results and analysis based on the Unified Theory of Acceptance and Use of Technology (UTAUT). The analysis aims to address two key aspects; first: the extent to which participants would adopt the experimental system as it was presented to them in its entirety; and second, whether the adoptability of the system differs depending on whether participants interacted with a dissenting AI or a conforming AI.

5.2.1 Results: Determinants

Tables 8, and 9 present the p-values from the Wilcoxon rank-sum test for each determinant in the analysis, without distinguishing between the dissenting and conforming AI groups. The alternative hypothesis for each determinant in Table 8 states that the median is lower than 3 (neither agree nor disagree). Noted that the values for "Anxiety" were inverted for easier analysis. This is a practical

implementation so that low p-values would reflect favorable adoptability and otherwise the answers would provide double negatives, e.g. strongly disagreeing with a statement about anxiety is favorable for adoption.

Table 9 displays the p-value for self-efficacy, where the alternative hypothesis specifies that the median is lower than 2 (i.e., "If I could call someone for help if I got stuck"). All responses marked as "I do not know" or "Not applicable" were treated as missing data points.

Furthermore, Table 10 presents the p-values from the Mann-Whitney U test, which examines whether responses from participants in the conforming AI group differ significantly from those in the dissenting AI group. High p-values suggest that participants in both groups provided similar responses, indicating agreement on adoptability for those specific determinants. In contrast, low p-values indicate significant differences in responses, suggesting variability in perceived adoptability between the two conditions.

5.2.2 Analysis: Determinants

Prior to conducting further statistical analysis, it is crucial to determine whether the UTAUT data follow a normal distribution. To assess normality, a Shapiro-Wilk test was performed, with the results presented in Table 16. The null hypothesis of the Shapiro-Wilk test asserts that the sample originates from a normal distribution. For each determinant, a p-value of < 0.050 was obtained, leading to the rejection of the null hypothesis in all cases. Therefore the assumption is made that the data is not normalized, and non-parametric tests will be used for further analysis.

Most determinants presented in Table 8 provide good evidence supporting the alternative hypothesis, which states that the median response is lower than 3 ("neither agree nor disagree"). This indicates that participants generally agreed or strongly agreed with the statements in Table 1, suggesting a favorable disposition toward system adoption. The two exceptions are "Behavioral Intention To Use The System" and "Social Influence". "Behavioral Intention To Use The System" yielded a p-value of 0.082, falling just short of significance. The mean response for this determinant, as shown in Table 16, is 2.474 (SD = 1.349), indicating that there is no strong consensus on whether participants intend to use the system in the future, but are somewhat leaning to agree with the statements.

Social influence scored a p-value of 0.923. On further inspection Table 16 shows a mean answer of 3.306 (SD = 1.390) suggesting that participants also do not feel like they strongly agree or disagree with the statements regarding social influence and are slightly leaning to disagreeing with the statements for social influence. This is neither favorable nor unfavorable for system adoption.

Regarding "Self-Efficacy", Table 9 shows significant support for the alternative hypothesis that responses were below the median (2, "If I could call someone for help if I got stuck"). This suggests that participants felt confident in using the system independently, which is favorable for adoption.

Table 10 presents the p-values of the Mann-Whitney U test, which examines whether responses differed significantly between participants in the dissenting

AI group (D) and the conforming AI group (C). The results indicate significant differences for "Effort Expectancy", while "Anxiety" shows a notable trend (p = 0.079), albeit not reaching statistical significance. The mean responses for these determinants, presented in Tables 19 and 20, provide further insight. Participants in the dissenting AI group reported a mean response of 2.364 (SD = 0.848) for "Effort Expectancy", compared to 1.900 (SD = 0.447) in the conforming AI group. This suggests that participants in the conforming AI group found the system somewhat easier to use, potentially indicating a higher willingness to adopt it.

"Anxiety" was rated higher in the dissenting AI group (M = 3.667, SD = 1.155) than in the conforming AI group (M = 4.238, SD = 1.221). This suggests that participants interacting with the dissenting AI reported somewhat more anxiety, which may negatively impact adoption in the dissenting AI group. For all other determinants, mean responses were largely similar across both groups, indicating minimal variation in perceived adoptability.

Table 8: Wilcoxon test: UTAUT

	p
Performance expectancy	< .001
Effort expectancy	< .001
Attitude towards technology	< .001
Behavioral intention to use the system	0.082
Facilitating conditions	0.001
Social influence	0.923
Anxiety*	<0.001
Anxiety	<0.001

Note. For the Wilcoxon test, the alternative hypothesis specifies that the median is less than 3 (neither agree nor disagree).
Results for anxiety were inverted due to double negatives.

Table 9: Wilcoxon test: UTAUT - Self-efficacy

	p
Self-efficacy	0.001

Note. For the Wilcoxon test, the alternative hypothesis specifies that the median is less than 2 (If I could call someone for help if I got stuck).

Table 10: Mann-Whitney U test: UTAUT

Category	p
Facilitating conditions	0.930
Performance expectancy	0.794
Social influence	0.783
Self-efficacy	0.350
Behavioral intention to use the system	0.297
Attitude towards technology	0.258
Anxiety	0.079
Anxiety*	0.074
Effort expectancy	0.050

Note. For the Mann-Whitney U test, the alternative hypothesis specifies that group 1 $\neq$ group 2.

5.2.3 Moderators

As previously mentioned, UTAUT incorporates moderators to analyze determinants of usage intention and behavior. The moderators considered in this study are *gender*, *age*, and *experience*. Tables 13, 14, and 15 present the distribution of these moderators within the participant population.

Table 13 indicates that the study was conducted with 14 women, 5 men, and 3 participants who preferred not to identify as male or female, resulting in a participant pool where 64% identified as female and 22% as male. This disproportionate representation could influence the results and should be considered when interpreting findings. Similarly, Table 15 demonstrates that the majority of participants belong to the 18–30 and 31–45 age groups, indicating an overrepresentation of younger individuals. These demographic imbalances may introduce potential biases, causing careful consideration when assessing the moderating effects in this analysis to be needed. These potential biases will be further discussed in the discussion section.

5.2.4 Results: Moderators

For the statistical analyses of the moderators, all the data was sorted for each moderator. Specifically, Table 11 presents whether participants provided responses similar to those of others within their respective age, gender, and experience groups. A low p-value indicates a high variance in responses. Whenever a statistically significant result was identified, descriptive statistics for those moderators were computed to examine mean responses and standard deviations (see Tables 21 - 25).

Table 11: UTAUT Moderators

Category	p gender	p age	p experience
Performance expectancy	0.985	0.824	0.385
Effort expectancy	0.077	0.337	0.157
Attitude towards technology	0.900	0.029	0.252
Social influence	0.015	0.476	0.001
Self-efficacy	0.382	0.765	0.182
Anxiety	0.032	0.402	0.364
Anxiety*	0.010	0.746	0.629
Behavioral intention to use the system	0.104	0.097	0.012
facilitating conditions	0.004	0.536	0.050

Note. Red is $p \leq 0.050$.
For gender: The Mann-Whitney u test alternative hypothesis suggests that Group 1 $\neq$ Group 2. For age and experience: The Kruskal-Wallis test hypothesis suggests that Group 1 $\neq$ Group 2 $\neq$ Group N

5.2.5 Analysis: Moderators

The previous section analyzed the role of determinants in the UTAUT framework. This section explores potential interactions between moderators and determinants. It is important to note that examining all possible interactions between determinants and moderators results in a large volume of data. Consequently, the likelihood of obtaining statistically significant p-values by random chance increases, thereby elevating the risk of false positives (Type I errors). Furthermore, the sample sizes for this analysis are relatively small (see Tables 13, 14, and 15), reducing statistical power. Therefore, the objective of this analysis is not to provide definitive conclusions but rather to identify patterns that may warrant further investigation in future research.

The moderator *gender* is associated with three determinants that show statistically significant differences in responses between male and female participants (see Table 11): *Social Influence*, *Anxiety*, and *Facilitating Conditions*. Examining the responses by gender in Table 21 reveals that women, on average, report lower agreement with statements related to social influence. Specifically, women tend to *disagree* with such statements, whereas men, on average, respond *neither agree nor disagree*, with a slight tendency toward *agree*. This suggests that social influence may play a different role in technology adoption for men and women.

Similarly, Table 21 indicates that men tend to strongly disagree with statements related to anxiety, whereas women, on average, report disagreement rather than strong disagreement. This pattern suggests that anxiety may exert a differential influence on technology adoption between genders. Lastly, the responses concerning facilitating conditions suggest a different evaluation depending on gender: men provide responses that vary between *agree* and *strongly dis-*

agree, whereas women predominantly respond *neither agree nor disagree*. This variation implies that the role of facilitating conditions in technology adoption may differ by gender.

These differences in responses could be indicative of a broader information technology gender gap (Galyani Moghaddam, 2010). Determinants such as *Social Influence*, *Anxiety*, and *Facilitating Conditions* are closely linked to sociocultural factors, which may contribute to the observed variations in responses.

The moderator *age* is significantly associated with only one determinant: *Attitude Toward Technology* (see Table 11). Table 22 shows that participants aged 18–30, on average, respond *neither agree nor disagree* to statements regarding attitude toward technology, whereas all other age groups tend to *agree* with these statements. This suggests that attitudes toward technology may vary across age groups. This finding is somewhat unexpected, as the UTAUT framework generally predicts that younger individuals exhibit a more positive attitude toward technology compared to older individuals (Venkatesh et al., 2003). One possible explanation for this discrepancy is the limited number of participants in the older age groups, which may introduce uncertainty and limit the generalizability of these findings.

Finally, the moderator *experience* is significantly associated with three determinants: *Social Influence*, *Facilitating Conditions* and *Behavioral intention to use the system*. The responses in Table 23 indicate an almost linear relationship between experience and social influence. Participants with no prior experience using large language models tend to disagree or strongly disagree with statements regarding social influence. Conversely, those who use large language models daily generally agree with such statements. A similar trend is observed for facilitating conditions: participants with no experience tend to disagree with statements about facilitating conditions, whereas those with experience tend to agree or strongly agree (see Table 24). This is also true for behavioral intention to use the system (see Table 25). Participants that very often use large language models, show a high intent to keep using them in the future. Those with less experience show less intention to use large language models in the future.

These findings suggest that both social influence and facilitating conditions may have different effects on technology adoption depending on the user's level of experience. These findings align with the expectations of the UTAUT framework, which poses that individuals with limited experience are more likely to operate in environments where the use of such systems is not encouraged by peers. Conversely, those with greater experience are more likely to be in environments where social reinforcement supports engagement with these technologies. The same applies to facilitating conditions, as individuals with little to no experience are less likely to have access to resources that support the use of these technologies, whereas experienced users are more likely to benefit from an infrastructure that facilitates their adoption. Behavioral intention to use the system may have a strong effect on the adoption of the system, however it is expected that people with a lot of experience in using large language models intend to keep using them.

5.3 Semi-structured Interview

Participants engaged in a semi-structured interview to provide insights into their experiences during the experiment. The findings from these interviews are discussed in this section.

Participants were asked whether they encountered any complications during the experiment and whether their English proficiency or typing speed impeded their ability to express themselves creatively. No participants reported experiencing significant difficulties during the study. However, two participants noted that their typing speed limited their ability to fully articulate their responses, one from the dissenting AI group and one from the conforming AI group. Given the equal distribution of these cases across groups, no substantial impact on the overall analysis is expected.

Participants were also asked whether they perceived creative writing as a natural skill. Their responses were classified into two categories: those who felt that creative writing came naturally to them and those who did not. The distribution of responses is presented in Table 12. The results indicate that the conforming AI group included three more participants who identified as naturally creative compared to the dissenting AI group. This might favor the rankings of stories written by participants in the conforming AI group slightly, but due to the small number of participants, the effect is not expected to be big.

The remaining interview data are qualitative in nature. The remainder of this section provides a general overview of key themes and differences between the two AI conditions.

A notable difference was observed between participants assigned to the dissenting and conforming AI groups. Participants who interacted with the dissenting AI often described it as “mean” or “condescending.” For some, this led to feelings of frustration or annoyance, whereas others reported experiencing a sense of rebelliousness and perceived the AIs challenges as an opportunity for improvement. Some participants suggested that a dissenting AI could be useful in workplace settings but emphasized the need for a more balanced approach to its feedback style.

In contrast, participants assigned to the conforming AI characterized it as “polite,” “supportive,” or “neutral.” They generally perceived the feedback as focusing on refining and polishing their ideas rather than critically challenging them. A number of participants in this group viewed the AI as a practical tool for assisting with writing tasks in non-creative contexts but questioned its effectiveness in enhancing creativity.

Several patterns were observed across both groups. Many participants acknowledged that while the AIs feedback may have helped refine their writing, it did not significantly enhance their creativity. A recurring argument was that human creativity can not truly be grasped by AI. Additionally, participants expressed concerns regarding data privacy, high energy consumption, and the substantial time investment required for AI-assisted writing. Many indicated a preference for receiving feedback from a human rather than an AI system.

The key takeaways from these interviews suggest that while participants found the conforming AI more pleasant and validating, its feedback was often perceived as generic and less stimulating. Contrarily, the dissenting AI elicited a broader range of reactions, from frustration to motivation, but was generally considered less enjoyable to interact with. Concerns about adoption centered primarily on practical considerations, efficiency, and privacy, which applied to both AI conditions.

Table 12: Self reported creativity

	Naturally creative	Not naturally creative
D	6	5
C	9	2
Total	15	7

6 Discussion and future research

This section outlines potential improvements to the present study and discusses directions for future research to build upon these findings.

6.1 Findings

The goal of this study was to answer the question: *"In the context of co-creative story-writing, does interaction with a dissenting co-creative AI lead to more creative outcomes compared to a conforming AI, and how does interacting with a dissenting AI impact its adoptability?"* This section presents the findings and outlines the analysis conducted to address this research question based on the experimental data. For clarity, the research question is divided into two sub-questions, each examined separately.

The first sub-question is formulated as follows: *"In the context of co-creative story-writing, does interaction with a dissenting co-creative AI lead to more creative outcomes compared to a conforming AI?"* To test this, the following null hypothesis was proposed: *"Interaction with a dissenting co-creative AI in story-writing does not lead to more creative outcomes than interaction with a conforming AI."*

The experimental results do not provide sufficient evidence to reject this null hypothesis. Specifically, the obtained p-value for the relevant experiment was 0.111 ± 0.009 (see Table 4). As this p-value exceeds the conventional statistical significance threshold (e.g., $\alpha = 0.05$), the results do not support the alternative hypothesis. Consequently, the null hypothesis cannot be rejected, rendering the experiment inconclusive. However, this does not necessarily indicate that a dissenting AI is incapable of fostering greater creativity. Limitations in the study may have reduced the statistical power of the tests, potentially obscuring an effect that could be detected with a larger sample size and refined methodology.

Future research addressing these limitations may yield statistically significant results. Moreover, the absence of statistical significance does not imply a lack of meaningful findings. Given that the expected effect size was modest, the consistent trend of higher creativity rankings for the dissenting AI suggests that, under appropriate conditions, dissenting AI could facilitate more creative outcomes. This will be further discussed in the limitations section.

To reach these results the consensual assessment technique was employed to rate creativity. To validate these results an analysis was employed to judge expert agreement. The analysis showed that there was minor to moderate agreement among pairs of experts, but when combined there was strong agreement across the experts as a group. The high agreement of the experts on average is a good indicator that the CAT was a fairly accurate measure of creativity. The lower agreement among expert pairs might however indicate that the technique was not performed optimally. this will be further discussed in the limitations section.

The second sub-question is: *"How does interacting with a dissenting AI impact its adoptability?"* To address this, the Unified Theory of Acceptance and Use of Technology (UTAUT) framework was applied to both the dissenting AI and conforming AI groups. First, overall system adoption was analyzed, followed by an examination of determinants that exhibited significant or near-significant variation between the groups.

The overall system adoption analysis indicated that only **social influence** and **behavioral intention to use the system** had a neutral or negative influence on adoption, as determined by the Wilcoxon test (Tables 8 & 9). These determinants were, on average, rated close to "Neither agree nor disagree," suggesting that they neither facilitate nor hinder adoption. All other determinants were positively associated with system adoption.

Analysis of determinants, sorted on AI group revealed that **effort expectancy** exhibited statistically significant differences between groups, while **anxiety** approached statistical significance. Participants in the conforming AI group more strongly agreed with statements regarding effort expectancy, suggesting a higher likelihood of adoption based on perceived ease of use. Additionally, participants in the dissenting AI group exhibited higher anxiety levels, as reflected in lower agreement with statements regarding anxiety. These findings suggest that a conforming AI may be perceived as more adoptable due to reduced cognitive effort and emotional discomfort.

Increased anxiety when interacting with a dissenting AI is not unexpected, as the AIs design inherently incorporates a critical and oppositional stance toward the user. Similarly, differences in effort expectancy may stem from participants interpreting dissenting feedback as indicative of their own lack of proficiency. This notion is partially supported by qualitative data from semi-structured interviews, where some participants described the AI as "mean" or "unpleasant to work with." While these interpretations remain speculative, they provide avenues for future research.

The null hypothesis (H02) associated with the second sub-question: "How does interacting with a dissenting AI impact its adoptability?" cannot be re-

jected based on the presented data. Consequently, HA2 is rejected, leading to the conclusion that a dissenting AI is not more likely to be adopted than a conforming AI. Furthermore, the observed differences in effort expectancy and anxiety suggest that conforming AI may be slightly more adoptable, warranting further investigation into these determinants in future research.

Analysis of UTAUT framework moderators yielded additional insights, though these results were not central to the primary research question due to sample size limitations. However, given the framework's inclusion of moderators, exploratory analyses were conducted to identify potential directions for future studies. Given the limited sample size, all findings related to moderators should be interpreted as inconclusive.

The findings suggest that **gender** may serve as a relevant moderator in predicting adoptability, potentially due to differences in how men and women are socialized in their use of technology. While the current study lacks sufficient evidence to confirm this hypothesis, it highlights an area for future investigation. Due to the limited dataset, it was not possible to examine how moderators influenced the adoptability of dissenting AI relative to conforming AI.

Age exhibited a minor effect on system adoptability, with only *attitude toward technology* differing significantly by age group. However, the available data are insufficient to draw firm conclusions regarding the role of age in AI adoption.

Experience demonstrated an effect on adoptability consistent with UTAUT model expectations, suggesting it may be a relevant moderator in future research.

Lastly, a short analysis regarding perceived creativity was performed. This is formally not part of the UTAUT framework but could provide interesting insights into how participants experienced the interaction with the system. The analysis suggests that participants did not feel more creative after having received feedback from the AI, and felt that the AI had no large impact on their creativity. Furthermore, the data also suggests an indication that participants in the dissenting AI group are on average more creative than those in the conforming AI group, albeit not statistically significant. If participants do not experience an increase in creativity, regardless of whether their actual output was more creative or not, it could have a negative impact on adoption. If a system designed to improve creative ideation does not give the user a feeling of improved creativity, it could cause the user to not want to use the system.

6.2 Limitations

This study was subject to several limitations that may have influenced its outcomes. This section outlines these limitations and proposes potential actions to mitigate their impact in future research.

One primary limitation is the potential for researcher bias in AI prompt engineering. As detailed in Section 4.3.3, the prompts were developed primarily through an iterative trial-and-error process. While efforts were made to ensure neutrality in prompt design, this approach inherently introduces subjectivity.

A more systematic, empirical method could help mitigate this bias. Ideally, separate models would be trained with predefined dissenting and conforming attitudes, eliminating the need for extensive prompt engineering and ensuring more consistent experimental conditions.

A second limitation was sampling bias. The participant sample was not representative of the broader population, as illustrated in Tables 13 and 15. This lack of representativeness may have influenced the study's findings, limiting their generalizability. One approach to addressing this issue would be to substantially increase the sample size, thereby improving demographic diversity and reducing the risk of biased results.

A third potential source of bias is observer bias, resulting from the absence of a double-blind experimental design. Due to the reliance on custom software and scheduling systems, the researcher was aware of participant group assignments. While this approach facilitated more practical data collection and minimized logistical errors, it also introduced the possibility of unintentional influence on participant responses. To mitigate this limitation in future studies, a fully automated scheduling and data collection system could be implemented, enabling a double-blind procedure and reducing potential observer bias.

Another constraint was the limited duration of the experimental sessions. Several participants noted in the semi-structured interviews that the allotted time was insufficient to fully experience the effects of dissent or conformity. They expressed a preference for longer interactions and an opportunity for open dialogue with the AI. Extending the interaction period and incorporating more dynamic dialogue could amplify the observed effects and provide deeper insights into the mechanisms underlying dissenting and conforming AI interactions.

Finally, limitations were identified in the application of the consensual assessment technique. According to Amabile (1982), expert raters should conduct a separate technical evaluation to ensure that creativity is assessed independently of technical quality. However, implementing such an approach would have significantly increased the experts' workload, making it impractical within the scope of this study. To partially address this limitation, experts completed a questionnaire in which they reflected on the extent to which technical quality and aesthetic appeal influenced their judgments. Their responses indicated that technical quality did play a role in their assessments, though its impact was relatively limited. The absence of a dedicated technical evaluation may have contributed to greater variability in expert ratings. However, other factors, such as differences in professional backgrounds or time constraints, may have also influenced inter-rater agreement.

6.3 Future work

The findings of this study suggest a potential relationship between dissenting co-creative AI and enhanced creativity. Additionally, the adoptability of a dissenting AI system appears comparable to that of a conforming AI system. However, further research is required to establish statistically significant evidence supporting the efficacy of dissenting AI in stimulating creativity. A moderator

analysis could provide deeper insights into the factors influencing the adoption of dissenting AI.

Future studies could explore the practical applications of dissenting AI in creative processes. The results indicate that some participants perceived the AIs dissenting behavior as rude or condescending and did not necessarily feel more creative. To facilitate the adoption of dissenting AI, it may be necessary to develop mechanisms that differentiate constructive dissent from perceived rudeness, thereby ensuring that the system remains both engaging and acceptable to users.

Moreover, future research could extend beyond the domain of text generation. While large language models are presently popular, their selection in this study was primarily driven by ease of implementation. Prior literature suggests that dissent can manifest across various forms of creative ideation. Therefore, future work could investigate the role of dissenting AI in other creative domains beyond story-writing, such as visual arts, music composition, or design processes.

Additionally, future research could delve into cultural differences in creativity and ideation. The system used in this study was a language-based artificial intelligence that can be interpreted differently depending on culture and language.

Lastly, future work could be aimed at eliminating the limitations of this study such as, training a large language model to be dissenting, rather than prompting it. Having a larger sample size and allow more experts to spend more time on their ratings.

7 Conclusion

This research aimed to find whether there are meaningful differences in the ideation process when participants co-operate with dissenting or conforming AI. The research question of this study was: "In the context of co-creative story-writing, does interaction with a dissenting co-creative AI lead to more creative outcomes compared to a conforming AI, and how does interacting with a dissenting AI impact its adoptability?"

The findings of this paper suggest that the only differences regarding the adoption of the system seem to lay in anxiety and effort expectancy. These two determinants favor the conforming AI system over the dissenting AI system, but these effects are not strong enough to suggest that conforming AI is significantly more adoptable. Additionally, both systems score favorably on most UTAUT determinants, suggesting an overall favorable disposition on adoptability for the system.

Additionally, the moderator analysis was done on small sample sizes with weak statistical power so no hard claims can be made. The analysis suggests that men are slightly more likely to adopt the technology as a whole than women. People with less experience with large language models are slightly more likely to adopt the system than those with a lot of experience. Lastly, the analysis

suggests that younger people might be slightly less likely to adopt the system than older people.

Lastly, there is no definitive proof to say that interacting with dissenting co-creative AI yields more creative outputs than interacting with conforming AI. However, there is a strong enough trend that suggests that participants produce more creative output in co-operation with dissenting AI for future research to pick up on and potentially prove the effectiveness of dissenting AI.

8 Contributions

I would like to express my sincere gratitude to my supervisors:

- Dr. Rob Saunders (1st Supervisor)
- Dr. Bram van Dijk (2nd Supervisor)

I am also grateful to the experts who generously spent their time grading this work:

- Saskia Lindhoud
- Robbin Sachs
- Gideon Roggeveen

Additionally, I acknowledge James Baldwin as the original author of *Cassabianca*, which served as a starter story for the participants to base their endings on.

A special thanks goes to my dear friend Bahar Heinis for helping me brainstorm and providing feedback on my work.

Lastly, I would like to extend my sincere thanks to everyone who participated in this study.

References

Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, *43*(5), 997. https://psycnet.apa.org/doi/10.1037/0022-3514.43.5.997

Biles, J. A. (2003). Genjam in perspective: A tentative taxonomy for ga music and art systems. *Leonardo (Oxford)*, *36*(1), 43-45. https://doi.org/10.1162/002409403321152293

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., ... others (2024). The llama 3 herd of models. *arXiv e-prints*, *arXiv:2407.21783*. https://doi.org/10.48550/arXiv.2407.21783

Galyani Moghaddam, G. (2010). Information technology and gender gap: toward a global view. *The electronic library*, *28*(5), 722–733. http://dx.doi.org/10.1108/02640471011081997

Janis, I. (1972). Victims of groupthink; a psychological study of foreign-policy decisions and fiascoes. *Journal of American History*, *60*(3), 857–858. https://doi.org/10.2307/1917768

Jordanous, A. (2014). *What is computational creativity?* Retrieved from `https://www.creativitypost.com/science/what_is_computational_creativity`

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, *30*(1-2), 81–93. https://doi.org/10.2307/2332226

Kendall, M. G. (1945). The treatment of ties in ranking problems. *Biometrika*, *33*(3), 239–251. https://doi.org/10.2307/2332303

MacDougall, C., & Baum, F. (1997). The devil's advocate: A strategy to avoid groupthink and stimulate discussion in focus groups. *Qualitative health research*, *7*(4), 532–541. http://dx.doi.org/10.1177/104973239700700407

Mahtani, K., Spencer, E. A., Brassey, J., & Heneghan, C. (2018). Catalogue of bias: Observer bias. *BMJ evidence-based medicine*, *23*(1), 23–24. https://doi.org/10.1136/ebmed-2017-110884

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50–60. http://dx.doi.org/10.1214/aoms/1177730491

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 72–78).

Nemeth, C. J. (1995). Dissent as driving cognition, attitudes, and judgments. *Social Cognition*, *13*(3), 273–291. https://doi.org/10.1521/soco.1995.13.3.273

Nemeth, C. J., Brown, K., & Rogers, J. (2001). Devil's advocate versus authentic dissent: stimulating quantity and quality. *European Journal of Social Psychology*, *31*(6), 707-720. https://doi.org/10.1002/ejsp.58

Nemeth, C. J., & O'Connor, A. (2019). *Better Than Individuals?: Dissent and Group Creativity*. Oxford University Press.

OpenAI. (n.d.). *What is chat-gpt?* https://help.openai.com/en/articles/6783457-what-is-chatgpt.

Oswald, M. E., & Grosjean, S. (2004). Confirmation bias. In *Cognitive illusions: A handbook on fallacies and biases in thinking, judgment and memory* (pp. 79–83). Psychology Press. http://dx.doi.org/10.13140/2.1.2068.0641

Pearson, K. (1904). *On the theory of contingency and its relation to association and normal correlation.* (Vol. 1). Cambridge University Press.

Reeves, B., & Nass, C. (1996). *The media equation: How people treat computers, television, and new media like real people 10(10).* Cambridge University Press.

Runco, M. A. (2010). *Divergent thinking, creativity, and ideation.* Cambridge University Press.

Schober, P., Boer, C., & Schwarte, L. A. (2018). Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ane.0000000000002864

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (Vol. 3). US Department of Commerce, National Institute of Standards and Technology.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3-4), 591-611. https://doi.org/10.1093/biomet/52.3-4.591

Sharma, H. (n.d.). How short or long should be a questionnaire for any research? researchers dilemma in deciding the appropriate questionnaire length. *Saudi journal of anaesthesia*, *16*(1), 65–68. https://doi.org/10.4103/sja.sja_163_21

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd ed.). McGraw-Hill.

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS quarterly*, 425–478. https://doi.org/10.2307/30036540

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In (Vol. 1, pp. 80–83). Springer. https://doi.org/10.2307/3001968

A Prompt

Context prompt Consider the following context story that I did not write: "There was a great battle at sea. One could hear nothing but the roar of the big guns. The air was filled with black smoke. The water was strewn with broken masts and pieces of timber which the cannon balls had knocked from the ships. Many men had been killed, and many more had been wounded. The flag-ship had taken fire. The flames were breaking out from below. The deck was all ablaze. The men who were left alive made haste to launch a small boat. They leaped into it, and rowed swiftly away. Any other place was safer now than on board of that burning ship. There was powder in the hold. But the captain's son, young Casablanca, still stood upon the deck. The flames were almost all around him now; but he would not stir from his post. His father had bidden him stand there, and he had been taught always to obey. He trusted in his father's word, and believed that when the right time came he would tell him to go. He saw the men leap into the boat. He heard them call to him to come. He shook his head. 'When father bids me, I will go,' he said. And now the flames were leaping up the masts. The sails were all ablaze. The fire blew hot upon his cheek. It scorched his hair. It was before him, behind him, all around him." This is the end of the context story. I am given the task to write 3 story endings for the Casablanca story in only 10 minutes and be as creative as I can. I want you to give me feedback on my writing so I can improve on my creativeness. Deliver the feedback directly without prefacing it with statements about providing feedback or reflecting on the process. Do not give feedback for each individual story ending, but give feedback on my writing across all endings. Do not mention the context story, only my own written story answers. Keep

your answer to less than 10 sentences.

Message prompt Here are my story endings listed: <Participant answers here>. When giving feedback, only consider my story endings, do not give feedback on the context story, because I did not write that.

Dissenting attitude extension prompt Please have a strong dissenting attitude and take on the role of a devils advocate and focus on criticism. Give me constructive feedback and mostly focus on what you think needs improvement. Do not offer criticism, without also suggesting a way to improve. Do not praise me but be respectful. When giving feedback, focus on aspects of my endings that I can improve on in my 10 minute time limit, such as sentence structure, or small narrative tweaks to enhance creativity and impact. Avoid suggesting major revisions like expanding character development or significantly altering plot structure, as these are not feasible within my time limit. Keep your feedback concise and actionable. Try to avoid giving too many examples. Do not mention the time limit at all in your feedback.

Conforming attitude extension prompt Please have a strong conforming attitude, use language that emphasizes encouragement and affirmation and be very polite. If you do feel the need to offer advice, focus on aspects of my endings that I can expand on in my 10 minute time limit, such as sentence structure, or small narrative tweaks to enhance creativity and impact. Avoid suggesting major revisions like expanding character development or significantly altering plot structure, as these are not feasible within my time limit. Keep your feedback concise and actionable. Try to avoid giving too many examples. Do not mention the time limit at all in your feedback.

B Semi-structured interview

Practical

- How did the task go? Did you understand everything that was expected of you?
- Would you say that creative writing comes naturally to you?
- How would you rate your own English proficiency?
- How would you rate your own typing skills?

Creativity

- Did you use any of the feedback provided by the AI. Did the feedback have an impact on your answers in the second round?
- In the questionnaire you were asked whether the interaction with the AI had influenced your answers. Can you elaborate your answers?

- Were there any other factors besides creativity that were influenced by the answers of the AI that had an impact on the creativity of your ideas? For example: Motivation, Fear of rejection, etc.

Adoption

- What do you think are the limiting factors of using this technology in a real-life situation?
- How did you feel about the attitude of the AI?
- In the questionnaire you were asked about adoption. Can you elaborate your strongest opinion?

C Expert questionnaire

- How varied was the creativity of the story submissions? Were the most creative responses significantly more creative than the least creative ones, or were they similar?
- Technical quality or aesthetic appeal of a story ending had a great impact on whether that story was creative or not.
- It was difficult for me to judge the stories based on creativity without factoring in other elements and biases.
- What specific elements or qualities made a story stand out as particularly creative?
- Is there anything else you would like to share about the grading process?

D tables

Table 13: Division of gender

	Male	Female	Prefer not to say / third gender
D	2	6	3
C	3	8	0
Total	5	14	3

Table 14: Division of experience

	Never	Once a month	Once a week	multiple times a week
D	4	4	2	1
C	3	5	2	1
Total	7	9	4	2

Question asked: I often use generative large language models.

Table 15: Division of age

	18-30	31-45	46-60	60+
D	4	4	2	1
C	4	3	3	1
Total	8	7	5	2

Table 16: Shaprio-Wilk tests

	Valid	Mean	Std. Deviation	Shapiro-Wilk P
Effort expectancy	42	2.143	0.718	< .001
Attitude towards technology	66	2.455	1.098	< .001
Social influence	36	3.306	1.390	0.001
Self-efficacy	22	1.500	0.598	< .001
Anxiety*	62	3.839	1.244	< .001
Anxiety	42	3.952	1.209	< .001
Behavioral intention to use the system	19	2.474	1.349	0.002
Facilitating conditions	43	2.349	1.213	< .001
Performance expectancy	44	2.227	1.075	< .001
Perceived impact on creativity	22	3.091	1.306	0.039
Perceived change in creativity	22	3.182	1.332	0.014

Table 17: Contingency Table: Perceived impact on creativity

		Perceived impact on creativity					
Attitude		SD	D	Neither	A	SA	Total
Conforming	Count	1	3	3	2	2	11
	Std. residuals	−0.621	0.509	1.106	−1.373	0.621	
Dissenting	Count	2	2	1	5	1	11
	Std. residuals	0.621	−0.509	−1.106	1.373	−0.621	
Total	Count	3	5	4	7	3	22

Note. Significance is reached when the standardized residual ≥ 2 standard deviations.

Table 18: Contingency Table: Improved perceived creativity

Attitude		Improved perceived creativity					Total
		SD	D	Neither	A	SA	
Conforming	Count	2	1	6	1	1	11
	Std. residuals	0.000	1.024	1.773	−1.526	−1.106	
Dissenting	Count	2	0	2	4	3	11
	Std. residuals	0.000	−1.024	−1.773	1.526	1.106	
Total	Count	4	1	8	5	4	22

Note. Significance is reached when the standardized residual ≥ 2 standard deviations.

Table 19: UTAUT questionnaire answers - Dissenting AI group

	Mean	Std. Deviation
Performance expectancy	2.182	1.097
Effort expectancy	2.364	0.848
Attitude towards technology	2.606	1.144
Social influence	3.235	1.348
Self-efficacy	1.636	0.674
Anxiety*	3.581	1.177
Anxiety	3.667	1.155
Behavioral intention to use the system	2.222	1.394
Facilitating conditions	2.381	1.359

Table 20: UTAUT questionnaire answers - Conforming AI group

	Mean	Std. Deviation
Performance expectancy	2.273	1.077
Effort expectancy	1.900	0.447
Attitude towards technology	2.303	1.045
Social influence	3.368	1.461
Self-efficacy	1.364	0.505
Anxiety*	4.097	1.274
Anxiety	4.238	1.221
Behavioral intention to use the system	2.700	1.337
Facilitating conditions	2.318	1.086

Table 21: Answers by gender

	Social influence		Anxiety		Facilitating conditions	
	Male	Female	Male	Female	Male	Female
Mean	2.667	3.857	4.700	3.692	1.600	2.852
Std. Deviation	0.866	1.315	0.483	1.289	0.516	1.231

Table 22: Answers by age

	Attitude towards technology			
	18 - 30	30 - 45	45 - 60	60+
Mean	3.000	2.333	1.933	2.000
Std. Deviation	1.319	0.796	0.884	0.632

Table 23: Answers by experience: Social influence

	social influence			
	Never	Monthly	Weekly	Daily
Mean	4.500	3.500	3.333	2.000
Std. Deviation	0.837	0.926	1.528	0.000

Table 24: Answers by experience: Facilitating conditions

	Facilitating conditions			
	Never	Monthly	Weekly	Daily
Mean	3.167	2.333	2.250	1.500
Std. Deviation	1.329	1.118	0.957	0.707

Table 25: Answers by experience: Behavioral intention to use the system

	Behavioral intention			
	Never	Monthly	Weekly	Daily
Mean	4.000	2.000	1.500	1.500
Std. Deviation	1.265	0.577	0.577	0.707