Norwegian School of Economics
Bergen, Autumn, 2023

Synthesizing Minds and Machines

An Empirical Study of the Impact of Human-AI Co-Creation on Creativity and the Moderating Role of Competence

Aleksander Stabell Eriksen & Anders Pay Eriksen

Supervisor: Eirik Sjåholm Knudsen

Master Thesis, Economics & Business Administration, Major in Business Analysis & Performance Management

NORWEGIAN SCHOOL OF ECONOMICS

Preface and Acknowledgements

This thesis was written as a part of the master's degree in Economics and Business Administration at the Norwegian School of Economics, where both authors major in Business Analysis and Performance Management (BUS). The thesis was written on behalf of Digital Innovation for Growth (DIG), Norway's leading research center on digital transformation and innovation for sustainable growth, and is thus a part of the DIGdeeper master program.

We wanted our thesis to explore and explain a topic at the frontier of innovation research. Hopefully, our findings will impact future decisions and affect how innovation processes are managed, contributing to management and decision-making in businesses. The work has been challenging and educational, but first and foremost, exciting and exhilarating. We express our gratitude to DIG for the opportunity to conduct our research and for granting us funds for data collection. Without DIG's contribution, our study would not be feasible.

Further, we wholeheartedly thank our energizing and always curious supervisor, Eirik Sjåholm Knudsen. Eirik has shown great interest in our project and provided unique ideas, solutions, and approaches. Your contributions and curiosity have been inspiring, motivating us to do our best. Thank you.

We also express our sincere gratitude to all the business leaders who agreed to act as judges in our experiment. Getting highly experienced professionals to evaluate the ideas was crucial for the credibility of our research. As we know how busy you are, we greatly appreciate the time and effort you spent to aid us with our thesis. Martin Schütt, Kjartan Kalstad, Jan Grønbech, Henrik Lisæth, Thomas Berglund, Per Christian Johannesen, Pål Brynsrud, Mads Agerup, Ingrid Lorange, Knut Glad, Pirasant Premraj, Klaus Røiri, Truls Braatas, Torkel Muri, Johan Karlsen, Stine Sofie Grindheim, Ole Larsen, Daniel Sørli, Ingrid Teigland, Johan Gjesdahl, Rolv-Erik Spilling, Gunnar Sellæg, Alex Adams, Torleif Ahlsand, Per Christian Tandberg, Kristian Næss, Alexander Haneng and Lars Flesland thank you.

Lastly, we thank our patient test participants Espen Møller Pettersen and Nikoline Aasnæs, and our third team member, the creative, always positive, supporting, and intelligent ChatGPT-4.

Abstract

This paper adds to the recent interest in the relationship between creativity and AI by studying if co-creating with AI positively affects creative outputs. Aiming to extend previous findings on human-AI co-creation as a creativity enhancer, and AI as a competence leveler, we conducted an experiment (n = 396) where students within knowledge fields generated business ideas, with or without ChatGPT-4. We find i) No difference in overall creativity between the groups co-creating with AI and non-AI users; however, optimal prompting produces significantly more novel ideas. ii) Competence does not moderate the relationship between AI and creativity when co-creating. iii) The non-AI group generated the two most creative ideas, while most of the top ten ideas were generated by prompted participants who co-create with AI. We conclude that AI can be a powerful creative tool, resulting in a potential interplay where AI is not a substitute for humans but a collaborator, amplifying human creativity and ingenuity.

Contents

List of Tables

List of Figures

1. Introduction

The latest advancements in generative AI challenge the assumption that creativity is a uniquely human ability and holds profound implications for knowledge workers in creative fields. AI systems, particularly Large Language Models (LLM), can now generate novel content virtually indistinguishable from human outputs across modalities (Jo, 2023). This leap from AI handling routine and automated tasks to engaging in intricate and creative activities signals a pivotal shift in the professional landscape. Given the potentially disruptive effect, the subject has naturally sparked researchers' interest, leading to a growing amount of research on whether AI or humans are more creative, with the results indicating that LLMs can match or outperform human creativity (Boussioux et al., 2023; Haase & Hanel, 2023).

While this research stream has made essential advancements in assessing AI's creative potential in isolation, there is a notable gap in research concerning how creative outcomes are shaped by the interaction between humans and AI in co-creation. Currently, most studies in AI and creativity focus on comparing the standalone creative outputs of various LLMs with those produced solely by humans (Guzik et al., 2023; Haase & Hanel, 2023; Boussioux et al., 2023), neglecting the more practical and common scenario where AI is employed collaboratively with human input. Consequently, while substantial knowledge has been gained about the creativity of AI on its own and the specific tasks where it excels independently, the potential impacts of collaborative human-AI co-creation is a nascent research field (Wan et al., 2023).

To contribute to the nascent research field, our thesis investigates whether human-AI co-creation positively affects creativity. Moreover, as recent research indicates that co-creating with AI can work as a leveler between highly competent and less competent workers (Dell'Acqua et al., 2023; Noy & Zhang, 2023), we further investigate whether competence moderates the relationship between AI and creativity. More specifically, our thesis will answer the following research question:

"Does human-AI co-creation positively affect creativity, and does the level of competence moderate the relationship?"

We conducted one experiment over two rounds with 651 students from knowledge-intensive fields to investigate the potential relationship. During the first round of the randomized experiment, the participants generated business ideas with or without AI in a pre-engineered chatbot (running on ChatGPT-4) to investigate whether human-AI co-creation positively affects creativity. Thus, the first experiment integrates two important streams of work: First, an established body of literature focusing on the creative output of different LLMs (Guzik, 2023; Haase & Hanel, 2023; Boussioux, 2023), and second, an emerging stream of research on the integration of human work with AIs through co-creation (Dell'Acqua et al., 2023; Noy & Chang, 2023; Choi & Schwarcz, 2023). In the second round, we provided all participants with AI access and guidelines on optimally using AI when co-creating to produce creative output. Hence, the second experiment investigates how to enhance the efficiency of creative co-creation and focuses on a small growing body of literature debating the optimal way of co-creating with AI in creative processes (Cope, 2005; Doshi & Hauser, 2023; Wan et al., 2023).

Our study has three main findings. First, the participants exposed to AI guidelines created significantly more novel ideas than those co-creating without AI guidelines and the non-AI users. However, there were no significant differences in overall creativity or idea usefulness. Our findings contrast with those of Boussioux et al. (2023), who suggested that AI increased usefulness and reduced the novelty of the ideas. Our findings indicate that the critical factor is not merely having access to AI but also knowing how to co-create efficiently with AI in creative work.

Second, our findings reveal that competence did not play a role in moderating the effects of AI, indicating that AI's impact on creativity is consistent across individuals, irrespective of their competence in the relevant field. This does not align with the conclusion of prior research (Dell'Acqua et al., 2023; Noy & Zhang, 2023; Choi & Schwarcz, 2023), which found a moderating effect for competence on quality, where the effect was most significant for workers of lower competence. However, we test for creativity and our student sample exhibits much less variation in competence than in a typical work-life setting, which might explain our findings.

Third, the most novel, useful, and overall creative ideas come from participants who did not use AI. This result is in line with previous research, where the best ideas came without AI intervention (Haase & Hanel, 2023), implying that for individuals with a high degree of creativity, AI involvement constrains their unbounded ideation rather than enhances it.

In conclusion, our study responds to the recent calls for more research on how co-creating with AI affects humans in creative endeavors (Cope, 2005; Haase & Hanel, 2023). Furthermore, our findings complement existing research on AI as a competence leveler (Dell'Acqua et al., 2023; Noy & Zhang, 2023; Choi & Schwarcz, 2023) and on whether AI or humans generate the most creative ideas (Boussioux et al., 2023; Girotra et al., 2023; Haase & Hanel, 2023). Most importantly, we contribute to research on the more practical scenario where AI is employed collaboratively in creative processes. This approach offers a perspective on how AI is integrated into creative processes in real-life settings rather than examining AI in isolation, as Guzik et al. (2023) and Girotra et al. (2023) did. Lastly, our findings provide insights into how knowledge workers can optimally leverage AI's capacity when co-creating creative work. Thus, the findings could give management and organizations some guiding principles for applying AI to optimally augment knowledge workers' creativity in practice.

2. Literature Review

2.1 Creativity

Creativity is a key driver for human advancement (Sternberg & Kaufman, 1999). It is crucial for problem-solving at work and in daily life; it drives new technologies, inventions, and scientific findings. Creative work also results in new products and services, improving life quality and creating jobs. There is no doubting the importance of creativity; however, what exactly is it?

Throughout the years, there have been several definitions and criteria for what creativity is and how to recognize it (Runco & Jaeger, 2012). Psychologists also agree that there are different creativity types and that different kinds of creativity require different skills and mindsets (Gaut, 2010). For example, Stein (1953) suggests that in physics, creativity means greater flexibility in the intellectual sphere, while in arts, it means greater flexibility in the emotional sphere. Moreover, Stein proposes a general definition of creativity, suggesting that creativity is novel work that proves useful or satisfying by a group at one point in time. Stein (1953), therefore, proposes two criteria for something to be considered creative. One, it must be novel, and two, it must be useful. By novel, Stein means something that did not previously exist in precisely the same form. Furthermore, Stein suggests that the extent to which work can be considered novel depends on how the work deviates from the status quo. With the phrase "useful or satisfying for a group…" Stein implies that some social judgment is needed to categorize a work as creative.

More recent research supports Stein's suggestion that creative work must be novel (Runco & Jaeger, 2012; Batey, 2012). However, the definition of "novel" varies somewhat, and the term has been used interchangeably with "original" in previous literature. Barron (1955) suggests that for something to be called original, it must, to some extent, be adaptive to reality, thus arguing that randomness and originality are not the same. Something completely random is likely original in the sense that it is uncommon and unique; however, to be original, it must also be of some use. Another way of describing novelty/originality is by requiring the creative work to have "effective surprise" (Bruner, 1962). Burner argues that novelty will surprise those exposed to it and, in extension, suggests that creativity partly can be measured by people's surprise when learning of an idea or a piece of work. He, therefore, considers the work to be original/novel if the work will surprise people. Nevertheless, as both Stein (1953) and Barron (1955) argue, novelty is not enough for something to be considered creative, though it is one criterion.

In addition to being novel, an idea must be worthwhile to pursue and represent some compelling property to be considered creative (Cropley, 1967). As Barron (1955) argued, randomness is not creativity; there must be some adaptation to reality. In other words, for an original thing to be creative, it must also be effective (Runco & Jaeger, 2012). Effectiveness takes many forms and has been given various labels by different researchers. In previous work, Runco (1988) stated that for creativity, "originality is vital, but must be balanced with fit and appropriateness." An idea that is fit and appropriate can likely be effective and useful. What Runco is stating is, therefore, that the work must be useful for someone at one point in time, supporting Stein's (1953) previous claims that both originality and usefulness are required for something to be creative.

Though Stein's claims have gained support, the phrasing and use of terminology vary. For example, Kneller (1965) states that a creative idea must be novel and relevant, while Cropley (1967) stated that the idea must be worthwhile to pursue. Though Kneller and Cropley use the terms "relevant" and "worthwhile" rather than useful or effective, most researchers seem to agree on the essence of a creative idea. It must be novel/original and be used in a sensible way that fits the problem the idea was to solve. Other researchers suggest quality as a criterion (Kaufman & Sternberg, 2007; Lubart & Guignard, 2004), emphasizing that for an idea to be creative, it must be novel and appropriate but also of sufficient quality. Whether the term used to describe the latter is effective, useful, adaptative, worthwhile, relevant, or sufficient quality is of little importance. The essence of the criterion remains the same.

As there are various opinions on the definition of creativity, concluding on a consensus definition seems unachievable, though a true consensus is nearly impossible in any field (Kaufman, 2016). Given the ambiguity of most terms associated with creativity definitions, settling on a simple, abstract, yet widely agreeable definition seems reasonable. Subsequently, for our purpose, a simple definition is sufficient. The two terms that reoccur most throughout the literature are novel and useful, and both terms are widely mentioned in most definitions (e.g., Batey, 2012; Mayer, 1999; Runco & Jae-ger, 2012; Zeng et al., 2009). Further on in our thesis, we will therefore consider the criteria for creativity to be novel and useful. How the two terms can be interpreted will be discussed when we review the literature on measuring creativity.

There are degrees of how novel or useful an idea is; in that regard, researchers distinguish between little C and big C creativity (Merrotsy, 2010). Little C creativity refers to ordinary people developing novel solutions to minor everyday problems (Simonton, 2017). Little C contains most of the creative pursuits of amateurs and hobbyists. For students, an example of little C might be making a new recipe and responding to some missing spices. On the other hand, big C creativity is not about recombining existing knowledge but about developing something truly new. Big C creativity refers to the production of ideas and innovation of importance, which is thus less intuitive and requires more time (Necka et al., 2006). More is needed than when creating novel solutions to everyday problems; the invention must shift the overall structure within a specific domain, such as inventing the wheel.

Measuring creativity is challenging if one does not measure a specific product or outcome of a creative process (Runco & Pritzker, 1999). Therefore, it is essential to know what should be measured and how. As we defined creativity as something novel and useful, those two criteria seem like the natural way to score something that aims to be creative. As stated above, most definitions have the same essence, though the exact wordings vary. When it comes to measurements, however, most models become more complicated. Cropley et al. (2011) created the CSDS, a 27-item scale used to measure creativity for management innovation. The scales comprised five main criteria: Relevance and effectiveness, problematization, propulsion, elegance, and genesis. Previous works have used other measurements, for example, unusualness, appropriateness, transformation, and condensation (Jackson & Messick, 1965). Though more criteria may give a more nuanced view of how something is creative, it is not necessarily essential when judging creative work.

Using the CAT method is a simple yet effective way of assessing product-based creativity (Said-Metwaly et al., 2017). CAT, short for the “Consensual Assessment Technique,” was introduced by Amabile (1982) and does not build on creativity theories. Instead, because of the ambiguous nature of creativity, CAT takes a more practical approach, evaluating ideas by letting experts in the relevant domain judge the products of creativity. Due to CAT’s similarities with how real-life creativity is judged, it has been referred to as the gold standard of creativity assessment (Baer & McKool, 2014) and has been frequently used to measure product-based creativity (e.g., Amabile, 1982; Christiaans, 2002; Kaufman et al., 2007). The approach supports Stein's (1953) previous statement that some social judgment is needed to categorize a work as creative and take a more realistic and practical approach to creativity assessment. Therefore, we will apply the CAT method when assessing the ideas generated by our participants.

Moreover, we must address the substantial research on intrinsic motivation’s effect on creativity. Though there is vast agreement that an individual’s motivation to perform a task is intrinsic or extrinsic (Deci & Ryan, 1985), there has been more uncertainty surrounding what motivates people to be creative. Nevertheless, research argues that people’s intrinsic motivation affects their creative output. Steiner (1965) states that to generate a creative solution to a problem, one must be motivated and inherently interested in the problem. Steiner’s research has later been elaborated and expanded on; Amabile (1983) found that creative outcomes depend on people’s intrinsic motivation, regardless of people’s creative capacity. These findings have gained support later. Prabhu et al. (2008) suggest intrinsic motivation mediates creativity, though to what extent somewhat depends on personality type.

Furthermore, a positive relationship was established by Fischer et al. (2019), supported by a previous meta-analysis suggesting a clear positive correlation between intrinsic motivation and creativity (De Jesus et al., 2013). Fischer et al. (2019) base their research on Amabile’s & Pratt’s (2016) creativity and innovation model, which emphasizes intrinsic motivation’s positive effect on creative outcomes. Taken together, the literature points to creativity being mediated and boosted by intrinsic motivation.

2.2 AI and Creativity

No 9000 computer has ever made a mistake or distorted information. We are all, by any practical definition of the words, foolproof and incapable of error.” The abovementioned quote is the fictional Artificial Intelligence computer Hal 9000, from the 1968 movie classic *2001: A Space Odyssey*, way of describing himself. Hal was eventually shut down by his co-pilot because of his aggressive and obsessive behavior towards real, though fictional, people.

There is probably some time until real AIs reach Hal's level; however, there is no reason to unplug the state-of-the-art AI models now. Contrarily, the past year has shown AI to be a powerful tool for those who know how to use it.

Though the evolution of AI can be considered a technological problem, applying AI is very much a problem for management. Therefore, researching how organizations and businesses can utilize AI in practice is crucial for gaining insight into the future of management, decision-making, and how to achieve sustainable growth. Disruption is coming – the Norwegian government has provided 1 billion kroner for AI research, taking one step closer to a future where AI is integrated with businesses and business models (Andreassen & Ugland, 2023). Further, millions of users made OpenAI’s generative AI ChatGPT the fastest-growing platform ever, making Microsoft invest 10 billion dollars in the company (Hu, 2023; Bass, 2023). Considering the heavy investments and rapid adoption, what makes the newest form of generative AI, Large language models (LLMs), so exciting?

LLMs are trained models predicting the next word dependent on an input text (openai.com). Some of the newer models are trained on billions of parameters through deep learning, making them adequate to respond to various questions and tasks across various domains. These models are called large language models and work as a probabilistic generative AI (Carlini et al., 2021). Simply put, the AI chooses what word is most likely to be next given the word beforehand, based on billions of human-written paragraphs (Wolfram, 2023). However, to ensure the text is not too predictable, the AI implements a level of randomness to its word selection. Therefore, the users will get different answers if they ask an LLM the same question multiple times.

The level of randomness can be adjusted by changing the LMM’s “temperature” (Wolfram, 2023) to make the LLM adapt and respond best to different questions; the temperature decides the probabilistic distribution of the words selected. A higher temperature will give the user more “random” words, resulting in more “creative” responses. In contrast, a temperature of zero will give the user only the most probable word to exceed the previously written text—the latter results in somewhat repetitive, deterministic, and confusing phrasings. For ChatGPT, a temperature around 1 (one) works best in most scenarios, and temperature = 1 is the default setting for the model (Carlini et al., 2021; Wolfram, 2023; openai.com).

Today, LLMs are few-shot learners, meaning they work well if the user provides a few examples of what content they generate (Brown et al., 2020). Unlike the earlier models, users can get meaningful responses without rigorously training the LLM on the specific task they want to complete. Current LLMs can also generate meaningful responses without training (zero-shot), where the user describes the task rather than providing specific examples.

Being trained as general models makes LLMs well-suited for responding to new questions and problems, as their general training has previously provided them with similar tasks to learn from (Dell´Acqua et al., 2023). The models are not simply generalists; however, ChatGPT-4 has proven to generate adequate responses when performing in professional settings within specific areas of expertise, such as law and medicine (Ali et al., 2022; Lee et al., 2023). Other models have proved to directly increase performance on specific tasks such as programming and writing, further supporting the belief that the new wave of AI is useful to knowledge workers (Peng et al., 2023; Noy & Zhang, 2023).

Subsequently, the most recent LLMs can possibly be of great use to knowledge workers. Previous disruptions in technology have mainly benefited low-skilled occupations by automating manual labor (Goldin & Katz, 1998). The same can be said of the previous AIs, such as machine-learning models, as they perform well at routine tasks that are easily codified and automated (Wang, 2019; Acemoglu & Restrepo, 2019). However, the recent release of LLMs has sparked a newfound interest in how AI may affect knowledge workers and aid in non-routine tasks (Dell´Acqua et al., 2023). As LLMs are generally trained systems, they can be of use in a variety of professions, and more and more research supports the belief that the models positively affect cognitive work (Eloundou et al., 2023; Felten et al., 2023; Dell´Acqua et al., 2023).

Though LLMs may positively impact aspects of work, the models should not be trusted blindly, as they can make plausible but incorrect points (Dell´Acqua et al., 2023). Given the previously described way LLMs work, the models have no regard for what is true or false – they generate a plausible continuation of the previous words (Wolfram, 2023). LLMs can, therefore, produce what is referred to as “hallucinations”; wrong information presented as truth (Ali et al., 2022; Dell´Acqua et al., 2023). Furthermore, LLMs perform poorly at certain tasks that a computer might expect to excel at, such as math and other quantitative problems.

As it may be challenging to predict what areas LLMs excel in and where they are close to useless, the current capabilities of LLMs can be regarded as a jagged frontier (Dell´Acqua et al., 2023). Therefore, further research on what areas the models are helpful and how to best use them is necessary.

As our thesis concerns the usage of generative AI in the form of LLMs, we will hereby refer to LLMs simply as AI. When referring to AI, we are, therefore, not referencing other types of generative AI, machine learning, or other systems that may be considered within the realm of artificial intelligence unless we explicitly say so. Therefore, all usage of “AI” refers to the state-of-the-art LLMs.

The current AIs’ completely disregard “truth,” and consequently, people must take what the chatbots write with a grain of salt (Dell´Acqua et al., 2023). Though AI's lack of understanding of what is real and false, right or wrong, feasible or fantasy, may seem like a weakness, it can also be an inherent strength. Disregarding what humans consider realistic or factual enables generative AI to go beyond what is evident today and break the bonds that limit creativity. Therefore, it comes as no surprise that although still nascent, the literature concerning AI capabilities indicates that AI excels at creative tasks (Boussioux et al., 2023; Haase & Hanel, 2023; Dell´Acqua et al., 2023; Guzik et al., 2023).

Creativity in AI is not new; decades-old AIs created new music, jokes, and paintings (Cope, 1991; Cohen, 1995; Binsted, 1996). Previous research provides clear evidence that the most recent AIs can be creative in some capacity (e.g., Girotra et al., 2023; Doshi & Hauser, 2023; Guzik et al., 2023). For AI to be adopted by humans, however, it must lead to better results than what humans can achieve in isolation. Subsequently, we review the literature and discuss whether AI can compete with and/or augment human creativity.

First, however, we discuss the philosophical aspect of AI and creativity by discussing whether AI can be considered truly creative. There seems to be broad agreement that AI is not creative like humans, mainly emphasizing differences in the creative process (Kirkpatrick, 2023; Haase & Hanel, 2023). When AI is creative, it either reorders information (data it has been trained on) or combines existing concepts into something humans have not thought of before (Kirkpatrick, 2023). Thus, AI cannot create purely novel concepts, though it is excellent at mimicking human creativity. Furthermore, Haase & Hanel (2023) stress that current AIs cannot trigger a creative process. Initiating creative processes is still unique to humans (Boussioux et al., 2023; Haase & Hanel, 2023; Dell´Acqua et al., 2023).

AI's output can still be novel in the sense that the output has not been seen before. Combining existing concepts into a new concept still creates something new. Subsequently, there is no arguing that AI can generate novel responses (Kirkpatrick, 2023). Various AI outputs have also proven useful and novel, making AI match our criteria for creative products (Dell´Acqua et al., 2023; Girotra et al., 2023). Therefore, retrieving and recombining knowledge is a perfectly fine form of creativity (Haase & Hanel, 2023). For practical instances, current AI systems can thus be considered creative as they generate novel and useful content. Whether this content creation process can be classified as truly creative or not may be up for debate, but we do not touch further on the subject in our thesis. As for our purposes, the literature indicates that AI is creative.

Current AI can match or outperform human creativity in certain domains (Girotra et al., 2023). ChatGPT-4 has proved to create better business ideas than students at Wharton Business School; when given the same prompt as the students, ChatGPT-4 produced better ideas on average and more ideas of great quality. Both zero-shot and few-shot prompting outperformed the students, though there was no significant difference between the zero-shot and few-shot prompted models.

Moreover, various chatbots, including ChatGPT-3.5 and ChatGPT-4, scored just as well as humans on the "Alternative Uses Test" (AUT), a frequently used creativity assessment test (Haase & Hanel, 2023). Notably, the most creative humans scored better than the best chatbots, though humans and AI performed equally on average. Other findings suggest that though AI can create novel and useful ideas, the level of novelty of purely AI-generated ideas is lower than that of humans (Boussioux et al., 2023). When comparing circular business solutions created by AI and humans, humans averaged higher novelty scores, while AI scored better at environmental and financial impact. The results differ somewhat from the findings of Girotra et al. (2023) and Haase & Hansel (2023), though it must be mentioned that the creativity measurements vary between the studies. Nevertheless, all studies suggest that state-of-the-art AI can match, if not outperform, human creativity, indicating that AI on its own is capable of generating creative solutions of sufficient quality on average.

However, organizations seek a few great ideas rather than an abundance of mediocre ones because they only have time and resources to pursue the very best (Girotra et al., 2010). Previously, we mainly discussed human vs AI creativity and the quality of their respective ideas on average. When performing repetitive tasks, like landing a plane or driving a bus, increasing the quality of the worst performance is desirable, and one should aim for a high average quality. If a 7/10 performance is adequate when landing a plane, an airline would much rather have all their landings be 7/10, as opposed to a half being 10/10 and half being 1/10. Unlike aviation, idea development is not a repetitive task (Girotra et al., 2010). Therefore, measuring the average idea quality is not necessarily appropriate when comparing human- and AI-generated ideas. Instead, Girotra et al. (2010) argue that when comparing ideas generated by different groups, only the best ideas from each group should be compared, as only the 10/10 ideas are relevant for further development.

Furthermore, the best way to generate a few great ideas may be to generate a large number of ideas (Girotra et al., 2023). In other words, quantity seems to affect quality positively. Given AI's quick content generation (Wolfram, 2023), one might expect AI's best ideas to be better than humans' best ideas. However, as previously discussed, results vary somewhat. Girotra et al. (2023) found that most of the best ideas from a sample of purely human-generated and purely AI-generated ideas were generated by AI. Other findings suggest that the most creative human ideas are more creative than AI's most creative ideas (Haase & Hanel, 2023). However, Boussioux et al. (2023) found no difference in overall creativity and quality among the best ideas, though their findings suggested that humans and AI differ in what aspect of creativity where they excel. Humans were more novel, while AI had more impactful ideas. Nonetheless, we stress that the three abovementioned studies applied different measurements for creativity and quality of ideas. The studies may have achieved similar results had they measured all constructs equally. Furthermore, all three studies compared purely human-generated to purely AI-generated ideas, while we aim to measure how AI affects human creativity through co-creation. Nevertheless, no matter how ideas are generated and measured, the best ideas should be compared to simulating ideation processes in practice.

Though AI may excel at creative behavior by generating creative content independently or augmenting human creativity, most studies agree that AI is less novel than purely human-generated content. Doshi & Hauser (2023) found that short stories co-written with AI were more enjoyable and creative. However, the AI-enabled stories were also more similar to each other than the purely human ones. Thus, the diversity of novel short stories was reduced when co-writing with AI, partly through writers anchoring on AI's similar ideas. Similar findings were produced by Boussioux et al. (2023), who found that creative business solutions created by humans had greater semantic diversity than those co-created with AI. The purely human-generated solutions also scored better on overall novelty. When Girotra et al. (2023) compared purely human-generated ideas to purely AI-generated ideas, the results remained the same: humans scored better on novelty, though AI scored better on overall creativity. However, Boussioux et al. (2023) argue that some of the novelty diversity in human-generated solutions may be perceived because individuals have different styles of presenting their solutions. In contrast, AI's style remains rigid, given the initial prompting.

2.3 Competence

Competence is often conceptualized as an individual's capability or knowledge in a particular domain (Norris, 1991). Competence embodies the synergy between skills, understanding of human interaction, and in-depth knowledge of a domain. Most scholars acknowledge its essential role in bridging the gap between educational frameworks and actual job-related skills (Boon & van der Klink, 2002; Mansfield, 2004). It also emphasizes the critical role of active learning and how individuals use learned concepts effectively.

There are three main approaches to competence, the first being "functional competence," the second being "cognitive competence," and the third being "social competence." First, functional competence is generally understood to be the combination of work-related skills, abilities, and in-depth knowledge within a domain (Nordhaug, 1993). A person with high functional competence acquires in-depth competence within a domain and demonstrates high performance to the standards required for employment in a work setting (Knasel & Meed, 1994; Dane, 2010).

Second, cognitive competence is more conceptual than the more operational functional competence. Cognitive competence is defined as knowledge, learning, and understanding that may be used in an occupational setting (Le Deist & Winterton., 2005). The term can be understood as conceptual competence and incorporates how we understand systems and recognize patterns, and apply them when analyzing information (Boyatzis, 2007). A cognitively competent person can solve problems by practically using learned concepts.

Third, social competence describes the ability to shape relationships and interact with others rationally and conscientiously (Le Deist & Winterton, 2005; Goleman, 1998). It embraces personality aspects such as our ability to cooperate and can be understood as our general behavior in interactions. A socially competent person uses emotional information about others that leads to or causes high performance (Boyantzis, 2007).

Most scholars identify a connection between the three defined competencies (Boyantis, 2007; Le Deist & Winterton., 2005; Mansfield, 2004). However, all three types of competence may not be relevant when individually generating ideas or handling business problem-solving. Although the level of social competence would predict the effectiveness of business ideas in a group setting (Mccallin. et al., 2007), social competence has hardly any impact when individuals generate ideas. When individually solving business cases, functional and cognitive competence give a more accurate prediction of how well the participants solved the business case (Carlsson & Eliasson, 1994). Therefore, the quality of generated ideas must refer to an individual's capability to use functional and cognitive competence in a domain-specific setting. The competence needed for business students could thus be described as business competence.

Students with a high degree of business competence manage to combine appropriate cognitive competence with technical skills (Tucker & McCarthy, 2001). Most business students are provided with an education that encompasses good decision-making skills related to working (King et al., 2001) and the ability to develop higher-order cognitive skills (HOCS) (Ennis, 1985; Zoller, 1999). Hence, one approach to measuring business competency could be testing a student's HOCS (Teijerio et al., 2013). HOCS can be understood as a form of cognitive competence and represents how an individual has acquired skills that could be used to make decisions and solve problems (Bradley et al., 2007). Furthermore, a higher degree of HOCS implies that a student has an improved ability to identify, integrate, evaluate, and relate concepts within a case study and can make the appropriate decision in each problem-solving situation (Hingorani & Sankar, 1998; Notar et al., 2002; Zoller, 2003).

A method of measuring HOCS is grade point average (GPA). Research indicates that GPA reflects the of students' intelligence, and as such, their cognitive competence (Bradley et al, 2007; Ickes et al., 1990). Bradley et al. (2007) examined how GPA affects students' perceived improvement in HOCS. The study suggests that as a student's cumulative GPA increases, the reported improvement in HOCS also increases. Hence, GPA could reasonably measure a student's HOCS, representing cognitive competence when handling business problem-solving.

Another approach to measure business competence could be to measure functional skills, as researchers believe that functional competence increases with the time spent in an occupation or a line of studying (Le Deist & Winterton, 2005; Drejer, 2000). According to Drejer (2000), an individual's level of functional competence develops from novice to expert if working on the same problem-solving tasks for a duration of time. As time passes, individuals learn to perform specific tasks better and better, meaning closer to the output objectives. We can assume that individuals develop functional working competence by repeating tasks and solving the same problems in an occupation (Dreyfus, 2004). While time spent in an occupation explains functional working competence, years of studying can explain functional educational competence. Hence, as a business student increases the number of years studying, the functional competence regarding solving specific business problems should increase.

2.4 Competence and Creativity

Competence and creativity are not opposing concepts but complementary aspects of idea generation (Campbell, 1960; Dane, 2010). Competence provides the foundational knowledge and skills, while creativity allows for applying this competence in novel and innovative ways. In that context, competence is necessary for creative idea generation (Dane, 2010).

Two different competence characteristics have been claimed to increase creativity. The first type is in-depth competence within a specific domain, which increases the complexity and knowledge depth within a domain (Dane, 2010). Deep knowledge in a specific domain will help individuals make more proficient use of their insights and previous experiences within a domain. Accessing their knowledge makes it easier for these individuals to identify and select promising linkages to domains for generating novel ideas. The second competence characteristic is broad domain competence, which broadens a person's knowledge base. How broad knowledge an individual possesses is represented by the number of domains within which an individual has some degree of competence (Dane, 2010). Broadening a person's knowledge base provides exposure to various domains, increasing their ability to recombine knowledge and create new linkages (De Dreu et al., 2008). As a result, the depth and breadth of competence affect creativity by shaping in-depth knowledge within one domain with a network of linkages to other domains.

The effects of these competence dimensions, however, have been mixed. On the one hand, individuals with in-depth knowledge within a domain have more complex knowledge structures and can thus consider more knowledge within the domain to create novel ideas (Amabile, 1996; Taylor & Greve, 2006). On the other hand, complex knowledge structures are also more prone to cognitive rigidity, which increases the inflexibility of linkages between domains and, therefore, limits individuals' ability to generate novel combinations (Audia & Goncalo, 2007; Dane, 2010; Mumford & Gustafson, 1988). When examining individuals with broad competence in different domains, they have greater flexibility to recombine knowledge across domains to generate novel ideas (Taylor & Greve, 2006; Simonton, 2009). However, flexibility can also cause new linkages that do not necessarily have a knowledge basis (Wadhwa & Kotha, 2006). Overall, this suggests that individuals need to have competence structures that are both complex and flexible to make novel ideas. They can access deep knowledge and make new novel linkages between domains when generating ideas.

Too much in-depth competence can lead to cognitive rigidity. Therefore, it is paramount that domain-specific competence is balanced by broad competence, especially as one becomes more specialized (Mannucci & Young, 2018; Dane, 2010). However, research indicates that in-depth domain competence is more important than broad competence when generating new, novel ideas (Mannucci & Young, 2018). In-depth domain competence acts as a sorting lens through which to consider new ideas. For example, individuals lacking in-depth domain competence may have problems distinguishing between genuinely novel ideas and ideas only new to them (Mannucci & Young, 2018). For economic students, this might indicate that generated business ideas become more creative and practical the longer they study. After a while, they will develop in-depth business competence, allowing them to sort out what ideas are truly novel and new.

2.5 AI Competence

Digital competence is the most overarching concept in describing technology-related skills (Ilomäki et al., 2011). The term is rapidly evolving as digital technology has reduced competency cycles: digital skills that were crucial a decade ago are no longer valuable (Gallardo-Gallardo & Collings, 2021). The newest addition to this competency cycle is the concept of AI competence, also labeled AI literacy.

The rapid growth in AI investment, which is expected to increase from $118 billion in 2022 to $300 billion by 2026, highlights the urgent need for AI skills (Shirer, 2022). Despite this, a significant gap exists between the current skill level and the skills needed to use AI effectively (Kandlhofer et al., 2016; Anton et al., 2020; Ångstrøm et al., 2023). To bridge this gap, the concept of 'AI literacy' has been introduced (Wang & Yuan, 2022; Long & Magerko, 2020). AI literacy involves identifying, utilizing, and evaluating AI technologies. Crucially, one does not need to be an expert in AI theory or development to be AI literate. This concept is similar to computer and digital literacy (Ala-Mutka, 2011), focusing on effective use rather than deep technical knowledge.

2.6 AI, Competence, and Creativity

There is extensive research on creativity and competence, and an increasing amount of literature on the relationships between creativity and AI. The three put together, however, are yet to be thoroughly investigated as of the time of writing. Nonetheless, previous findings concerning AI and competence enable us to theorize about the relationship.

Felten et al. (2023) research indicates how AI influences different occupations, requiring different sets of competencies. They link the most prominent AI applications to 52 human abilities and over 800 occupations in the US. The exposed occupations are characterized as having mainly highly educated and paid white-collar workers. These findings are consistent with similar findings concerning newer AI (Eloundou et al., 2023) and other technological advancements (Benešová & Tupa, 2017). However, the results contradict findings in similar evaluations of overall exposure to machine learning (Brynjolfsson et al., 2018). While machine learning substantially affects manual work where functional competence is most important (Brynjolfsson et al., 2018; Manyika et al., 2017), AI seems to affect occupations needing more cognitive competence. As mentioned, cognitive competence is conceptual competence that incorporates how we understand systems, recognize patterns, and apply information when analyzing. AI excels in recognizing patterns and applying information; therefore, it makes sense that occupations mainly reliant on cognitive competence are the most affected.

Furthermore, several studies show that AI affects white-collar workers' performance differently depending on competence levels. A study conducted by Noy and Zhang (2023) provides evidence of how productivity among college-educated professionals is affected by ChatGPT. In their study, both poor and well-performing professionals perform better using ChatGPT. However, while the well-performing (50% best) professionals have slightly increased quality and productivity, the poor-performing (50% worst) have a considerable increase in both (Noy & Zhang., 2023). The findings imply that workers with lower domain-specific competence benefit more from AI. The findings are consistent with the findings of Dell'Acqua et al. (2023), where poor-performing management consultants saw the most considerable competence enhancement when solving business problems using ChatGPT-4. As creativity is essential when solving business problems (Reiter-Palmon & Illies, 2004), the findings indicate how AI affects creativity for people with high business competence in their specific field. The study also suggests that AI is a great leveler of competence for highly educated individuals. In both these studies, however, they only examined highly educated professionals who already somewhat excel within their occupation. Scholars have addressed the need for more research with individuals having more significant variations in competence, such as students (Choi & Schwarcz, 2023).

However, AI is also a leveler for students; ChatGPT-4 mitigated the inequalities between top and bottom-of-class law students (Choi & Schwarcz, 2023). Taking together with the findings of Dell'Acqua et al. (2023), these results indicate competence may moderate a potential relationship between AI and creativity, causing a stronger positive relationship for less competent workers. Our reasoning is further underpinned by the previously mentioned findings of Doshi and Howser (2023), who found short stories written with AI access more enjoyable, especially for less creative writers.

2.7 Hypotheses Development

Though AI seems able to match human output with initial prompting only, the best results may be achieved using AI as an interactive tool (Wu et al., 2021; Anantrasirichai & Bull, 2022). Interrogating the AI to avoid blindly adopting hallucinations or simply bad output is crucial, especially when operating outside AI’s current frontier (Lebovitz et al., 2022; Dell’Acqua et al., 2023). When operating inside the frontier, however, AI seemingly leads to better results, regardless of human interference, as suggested by a field experiment using highly trained knowledge workers from Boston Consulting Group (Dell’Acqua et al., 2023). Granted, neither Dell’Acqua et al. (2023), Dell’Acqua (2022), nor Lebovitz et al. (2022) tested for creativity in isolation but researched how AI affected overall performance for knowledge workers.

Human and AI co-creation has led to increased quality of creative outputs across a diverse set of creative fields (Miller, 2019). Doshi & Hauser (2023) provide evidence that short stories written with the aid of AI are more enjoyable and that AI increases the writer’s creativity. When writing short stories, AI can create possible starting points for the author to develop into coherent stories, offer suggestions for plot points, or overcome writer's block by creating the next step in the story. Furthermore, Wan et al. (2023) writers became more novel and creative, drawing inspiration from the AI’s unexpected and random output. Even failures, such as useless ideas, were cherished as seeds of inspiration. The results indicate more creative and better products through human-AI co-creation. Doshi’s & Hauser’s (2023) findings are not unique; Jia et al. (2023) found similar effects of AI augmenting human creativity in a practical work setting. Furthermore, Anantrasirichai & Bull (2022) reviewed AI technology and applications in the context of creative industries, concluding that AI’s design promotes the models to augment rather than replace human creativity.

Taken together, previous research suggests that human-AI co-creation may increase human creativity, generating more creative output than humans and AI are capable of on their own. Specific effects may be hard to agree on due to a lack of universally used measurement of constructs. However, several studies indicate that a positive relationship exists between the variables. We investigate the relationship further, introducing our first hypothesis:

H1: Human-AI co-creation positively affects creativity.

Although co-piloting with AI may increase creativity, state-of-the-art AIs work best when the user treats it as a person (Mollick, 2023), and the right prompting plays a significant role in AIs' creative output (Boussioux et al., 2023). Previous studies suggest AI performs better at general tasks when interacting with it through thorough iteration processes, almost as one would with a coworker (Lebovitz et al., 2022; Dell'Acqua, 2022). Further, more recent studies indicate that the same effect applies to creativity; the most creative outcomes are possibly generated with numerous interactions (Doshi & Hauser, 2023; Wan et al., 2023). Wan et al. (2023) suggest that co-creating with AI in creative processes becomes most effective when the creative process is iterative. Writers became more novel and creative, drawing inspiration from the AI's unexpected and random output. Even failures, such as useless ideas, were cherished as seeds of inspiration. Similar observations are found for Doshi & Hauser (2023), where writers who interact more and receive more ideas in the creative process become significantly more novel and creative. Given AI's capability to generate vast amounts of ideas, the studies indicate that creative knowledge workers should seek a wide range of input when co-creating with AI, especially in initial ideation. These findings illustrate AI's role as a valuable collaborator when augmenting ideas or making implausible concepts more viable. Taken together with previous findings suggesting prompting plays a significant role in AIs' creative output (Boussioux et al., 2023), we want to investigate further how AI-user interactions affect the creative output of the co-creation. Subsequently, we introduce our second hypothesis:

H2: Human-AI co-creation produces more creative ideas when the user efficiently prompts and interacts with the AI.

As AI could be a competence leveler, and domain-specific competence seems to affect creativity positively, it is interesting to examine what effects domain-specific competence has on the relationship between AI and creativity. As discussed in the previous section, broad competence increases an individual's ability to combine existing concepts into new ones (De Dreu et al., 2008), while Manucci (2016) found that in-depth domain competence is even more critical when generating novel ideas. Such findings stipulate a positive relationship between in-depth competence and creativity. These are not moderating effects; however, Dell'Acqua et al. (2023) and Noy & Chang (2023) found a moderating effect of domain-specific competence on the relationship between AI and output quality. In the study, workers with lower competence saw the most considerable quality enhancement when co-piloting with AI. The same seems to apply to students as Choi & Schwartz (2023) found ChatGPT-4 to mitigate the inequalities between top-of-class and bottom-of-class law students. However, these studies examine the effects on outcome quality and not creativity.

Based on recent findings concerning AI's effect on performance differences, we theorize that domain-specific competence moderates the potential relationship between AI and creativity. Therefore, we test for competence as a moderator, not an independent variable affecting the proposed relationship between AI and creativity. We examine domain-specific competence through participants' GPA as a measurement of their cognitive competence and years studying within the current field as a measure of functional competence. Subsequently, our third hypothesis reads as follows:

H3: AI affects creativity differently, depending on the individual's domain-specific competence, regardless of how AI is used.

3. Methods

3.1 Research Model

Based on the reviewed literature, we developed the following conceptual research model to visualize our research question (Figure 1). Our study aims to investigate and explain the relationship between AI and creativity. The dependent variable (DV) is creativity, and the independent variable (IV) is AI. Moreover, we suggest that competence moderates the potential relationship between the two variables, thus introducing competence as a moderator variable (M). Our model can, therefore, be considered a moderation model, as it stipulates a third variable's moderation effect on the potential relationship between the independent and dependent variables (Fairchild & MacKinnon, 2009).

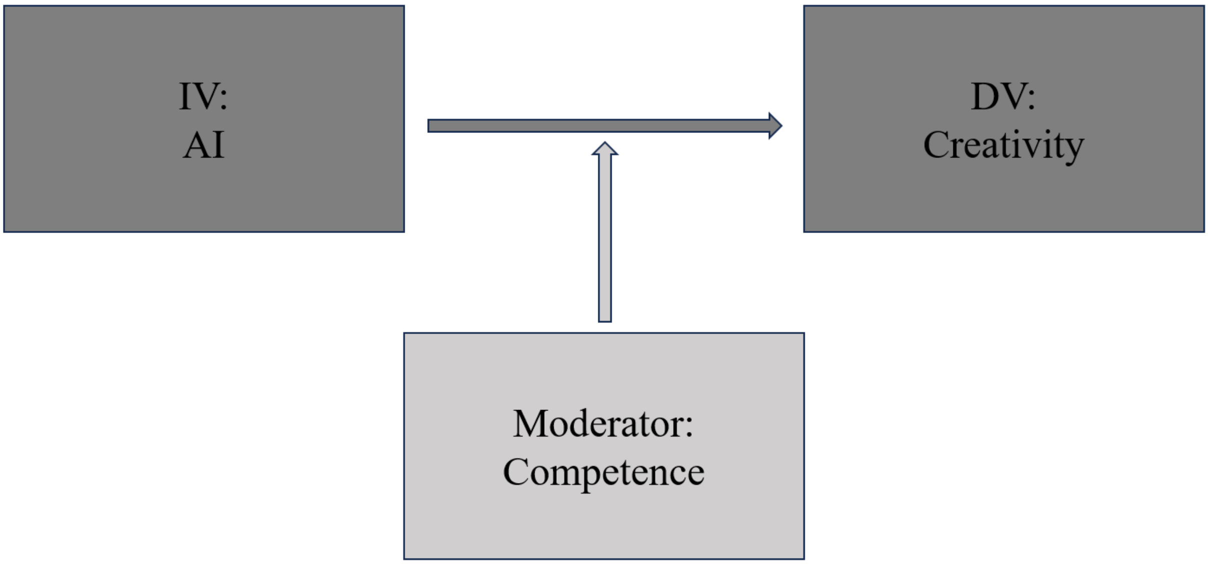


Figure 1 - Conceptual research model

Though the current literature points to AI augmenting and/or matching human creativity, findings are somewhat contradicted concerning human-AI co-creation and if and how people should utilize AI for creative purposes (Boussioux et al., 2023; Girotra et al., 2023; Haase & Hanel, 2023). We aim to enhance the current understanding of AI's uses and flaws by further investigating the potential relationship between AI and creativity. Moreover, we contribute to the existing research field by examining potential moderating effects to better understand the potential AI-creativity relationship. More specifically, we theorize that a person's competence on a topic moderates how AI may affect that person's creativity within that specific domain.

We test for this effect by measuring the creativity of business ideas generated by Norwegian students.

3.2 Research Design

Our research design explains our approach and strategy to answer the research question at the beginning of the thesis (Saunders et al., 2019). We explain our approach to theorizing and outline our chosen research strategy.

We took a clear theoretical position when theorizing and drafted our research question based on a thorough literature review. Based on previous findings, we generated and tested hypotheses by collecting and analyzing data. Thus, we apply a deductive research approach aiming to establish a causal relationship between two variables and investigate potential moderating effects (Saunders et al., 2019). When aiming to establish causal relationships, the research design can be considered explanatory as we try to explain potential variations in the independent variable.

A quantitative research strategy is appropriate because we want to test for causal relationships. A quantitative design choice facilitates numerical measurements of the variables, enabling regression analyses to test for causal effects (Saunders et al., 2019). We therefore consider an experiment as a natural choice of strategy, as it allows for causal hypothesis testing. Thus, an experiment applies to our research goal of establishing a causal relationship between AI and creativity and explaining what may cause the potential relationship.

The experiment was conducted online. We did so for two reasons. First, when running regression analyses, a sample size of a minimum of 50, plus eight for each predictor variable, is recommended (Green, 1991). Furthermore, research on moderation effects is particularly difficult when using small sample sizes (Fairchild & MacKinnon, 2009). Therefore, to ensure a sufficient number of participants, we opted for an online experiment that could be completed anywhere and within an extensive timeframe, lowering the threshold for participation. Second, as students, we have limited time and resources. Conducting an online experiment saved us substantial time as we did not have to supervise all participants.

Moreover, we cannot access suitable premises to create an environment sufficient for a lab experiment.

3.3 Data Collection

3.3.1 Sampling Procedure

Our experiment used Norwegian students as participants. Using students gave us a relatively homogenous sample, reducing the risk of other variables causing minor differences in the independent variable and allowing us to explain variations in creativity in greater depth (Saunders et al., 2019). Students are of similar age, life experience, and financial situation, giving them an equal foundation for ideation and problem-solving than people in entirely different demographic groups. Moreover, students often have a flexible schedule, making participation more accessible.

Though most participants attend the Norwegian School of Economics (NHH), participation was open to Norwegian students at all universities and faculties. However, we mainly targeted NHH students for two reasons. First, we are NHH students ourselves. Distributing the online experiment through known channels using NHH communication infrastructure and to our social network at school made contacting many students accessible. Second, we test for specific competence as a moderator variable. As the ideas generated were business ideas, business competence is the domain-specific competence relevant to our study (Norris, 1991). NHH students provide a set of people with low to high business competence, dependent on semesters completed at school and cumulative GPA (Le Deist & Winterton, 2005; Bradley et al., 2007). Therefore, NHH students are suitable when measuring how different levels of business competence moderate the effect AI may have on business idea generation. To compare results to people with little to no business competence, however, it was desirable also to include students within other fields. Subsequently, we distributed the experiment to students at the University of Bergen (UiB) and the Norwegian School of Science and Technology (NTNU).

The experiment was distributed through two channels: email and social media. At NHH, the experiment was first distributed to our social network through various social media channels, allowing us to get a preview of the data. Later, the experiment was emailed to all Norwegian students attending NHH. When emailing the experiment, we split the students into two groups, equally distributed across school years, in case we wanted to adjust or change the treatment. Two days after the initial distribution, all students were sent a reminder. To increase the sample size further, we promoted the experiment in popular lectures at the bachelor level, as these lectures typically achieve high attendance. We also distributed the experiment at UiB at the faculties for law and medicine and NTNU. At these universities, the experiment was distributed through social media channels by people in our social network studying at the respective universities.

We used volunteer sampling as all who received the experiment chose whether to participate (Saunders et al., 2019). Thus, we applied a non-probability sample mainly using self-selection. As we did not restrict participants from forwarding the experiment to friends or associates, there may have been snowball effects on our sampling.

To incentivize participation, the participants could enter a pool eligible to win four gift cards with a total value of 20,000kr. The price consisted of two gift cards at Norrøna of 5000kr each and two gift cards at DB Journey, also of 5000kr each. When we had concluded the data collection, four participants were selected using a random number generator and emailed their respective gift cards.

3.3.2 The Survey

The online experiment was developed in Qualtrics and represented a survey. Qualtrics offers an easy way to gather data and design online experiments without using code (Qualtrics.com). When entering the experiment, the participants were first subject to information concerning the experimental task, participation, and handling of personal data. After confirming they understood the task, the participants were divided into two groups by random assignment: one treatment group and one Control Group. Random assignment ensures both groups are equal except for the planned intervention manipulating the treatment group (Saunders et al., 2019). Thus, the possible external effects of an alternate explanation for variation in creativity are reduced. The whole survey design is illustrated in Figure 2.

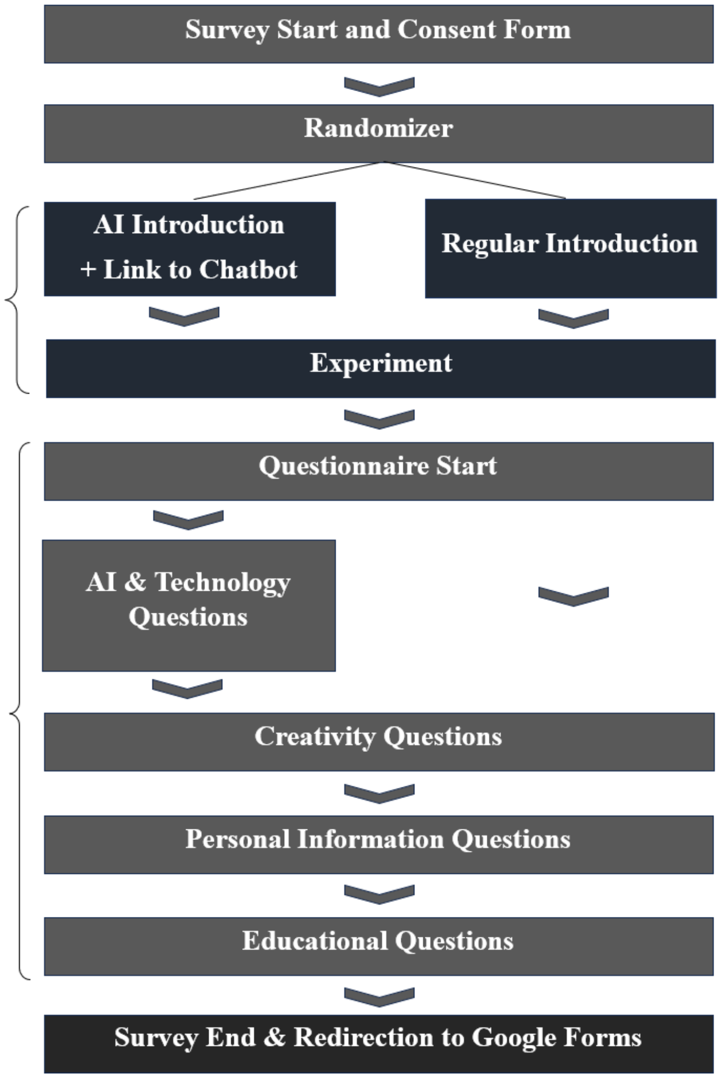


Figure 2 - Survey design, Survey 1

Both groups were assigned to generate a business idea directed at Norwegian students and were given 15 minutes to complete the task. The idea was directed at Norwegian students to ensure the participants had adequate knowledge of the target group and because narrowing down the potential ideas would make them easier to compare and score. The price should be low enough for the product to be accessible to the average student, but there were no other limitations concerning the product or revenue model. Participants could only register ideas after the two-minute mark to ensure all responses were given sufficient effort. The business idea had to be communicated through text and be a minimum of 10 and a maximum of 850 characters, approximately 150 words or one paragraph. Both groups were told to avoid communicating with others and to avoid external aids like the Internet. For the exact prompt given to participants and the experiment-participant interface, see Appendix 8.1.

The treatment group was given a link to a chatbot running on the state-of-the-art generative AI model ChatGPT-4. The chatbot was built using a chatbot builder named GPTtrainer, which allowed us to gain data from each specific interaction, make guidelines, and change the temperature of the responses (gpt-trainer.com; Boussioux et al., 2023). To simulate a normal conversation with ChatGPT4, we used the standard temperature. The chatbot was only available for participants in the treatment group, and subsequently, the AI access represented a planned intervention manipulating how the treatment group performed their assignment (Saunders et al., 2019). The prompt and temperature setting are displayed in Appendix 8.2. When interacting with the chatbot, the AI would introduce itself by saying it is available to help generate a business idea, encouraging the participants to interact and ask questions, as illustrated in Appendix 8.2. Note that we investigate how AI affects human-generated creative outputs, not whether AI is more creative than humans.

We altered the prompt of our chatbot slightly after the first interactions revealed a clear pattern. If the participants were to input a direct copy of the participant prompt, they all received the idea "StudyBuddy." Therefore, we changed the chatbot's guidelines to respond with various business ideas, not "StudyBuddy" as a generic response. However, all other prompts or messages given to the chatbot would have no limitations, simulating a normal conversation with ChatGPT-4. The changes were made after only three participants had interacted with the initial chatbot. Further, there were no differences between these three ideas and ideas generated with the altered chatbot, except for all three ideas being the same idea: StudyBuddy. Therefore, we chose to include the three ideas in our final sample.

A preview of the data collected from the initial distribution through our social networks revealed that most participants in the treatment group generally had only one or two interactions with the chatbot. Though an exciting finding, few interactions mainly measure "purely" AI creativity, not the desired output of human-AI co-creation. Literature suggested that co-creating with AI performs better when interacting through an iterating process (Lebovitz et al., 2022;) and that this also applies to creativity (Doshi & Hauser, 2023; Wan et al., 2023). Initially, we thought of having a second treatment that nudged participants to interact more and have a more iterative creative process. We decided to wait to introduce the second treatment, as we wanted to secure enough participants for our first hypothesis. However, a few days after launching the first experiment, we already had sufficient participants to test for differences between the Control Group and the Treatment Group. Therefore, we created a second treatment to test for the effects of efficient prompting and the nature of the interactions described through our second hypothesis. The treatment received the same questionnaire but had some adjustments in the introduction to the experiment. In addition to receiving access to AI, they were encouraged to interact with the chatbot and received five guidelines on how to efficiently prompt and interact with AI for creative purposes before the experiment. The survey design for the second survey is illustrated in Figure 3.

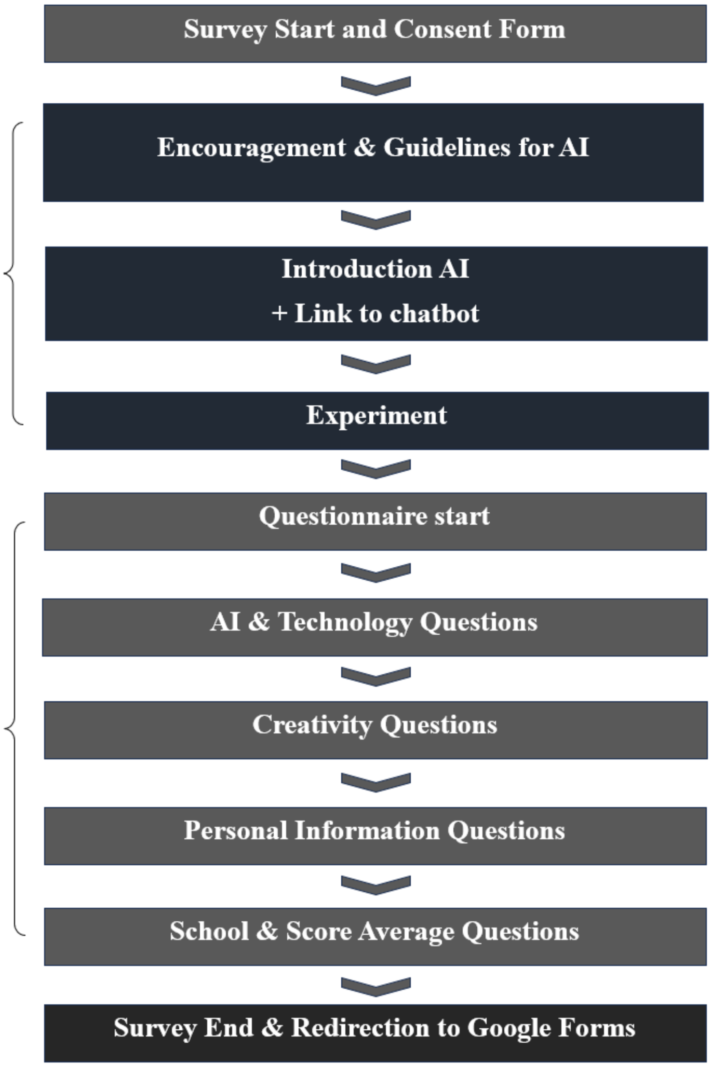


Figure 3 - Survey design, Survey 2

When distributing the experiment by email, 2/3 of the students were sent the original experiment with a 50/50 randomizer, placing participants in either the Control Group or the first treatment group. Furthermore, 1/3 were sent a new Qualtrics survey with the second treatment only. The survey the students received and the experiment was chosen randomly, ensuring we applied random assignment (Saunders et al., 2019). The final distribution of participants in the Control Group, Treatment Group 1, and Treatment Group 2 is shown in Table 1.

	Control Group	**Treatment Group 1**	**Treatment Group 2**
Experimental Task	Generating a business idea	Generating a business idea	Generating a business idea
Treatment	None	AI access	AI Access & guidelines on how to co-create with AI

Table 1 - Participant groups

In addition to collecting data from the experiment, the surveys included a questionnaire for gathering information about each participant. To prevent potential priming effects, the questionnaire was given after participants had completed the experimental task (Van den Bussche et al., 2009). Ideally, we would separate the experiment and the questionnaire to reduce the time spent working in one setting, thus reducing the probability of participants hurrying through to make other commitments (Saunders et al., 2019). However, we considered it unrealistic to make students participate in two settings and therefore opted to combine the experiment and questionnaire in one survey.

Participants in Treatment Groups 1 and 2 were asked questions concerning generative AI. These questions were not asked of the Control Group. The questions included when the participants used generative AI for the first time, how experienced they were at using AI, and their willingness to adopt new technologies. The questions were based on Venkatesh (2000) and Vărzaru (2022).

All participants were asked questions concerning the personality trait of openness to new experiences, which has proven to predict creative behavior (Harris, 2004). These questions consisted of five items from the Big Five personality test and were meant to measure the participants' self-reported creativity.

Moreover, we gathered basic information about the participants, such as age and gender. We also asked for their field of study, which was separated into six subgroups based on study fields provided by Utdanningsdirektoratet (utdanning.no, 2023). Those who studied economics and business administration were given an additional question concerning which school they attended. The additional questions allowed us to measure the number of participants studying at NHH and compare results between business schools.

Participants were asked for their GPA and years spent on their current degree to investigate competence. Bachelor students reported their GPA at the bachelor's level, and master students reported their GPA at the master’s level. Consequently, participants had to report at what level they were currently studying. Master students did not report their cumulative GPA for all their studies, as we are interested in their competence as of today. We, therefore, consider older grades irrelevant. Further, questions were asked concerning the years spent studying in the current school and whether the participants had professional experience within business.

Descriptive questions about age, gender, GPA, field of study, years spent studying, and professional business experience were placed last in the questionnaire. This was done to avoid priming effects like identifying oneself with an industry or cognitive capacity from affecting the self-reported values of the previous questions concerning AI experience and creative behavior (Meyer & Schvaneveldt, 1971). For the whole questionnaire, see Appendix 8.1.

When the survey was completed, participants would be redirected to a Google Forms, where they could assign their email address to the pool eligible to win the four gift cards. By collecting the emails through a different API, we could collect emails for the price lottery without connecting specific emails to specific responses. Thus, all responses remained anonymous in line with guidelines from FEK (De Nasjonale Forskningsetiske Komitéene) (forskningsetikk.no, 2019).

3.3.3 Preparation of the Data

Initially, 651 survey responses between 23/10/2023 and 2/11/2023 were downloaded from Qualtrics and divided into one survey with 421 responses and one with 229 responses. Furthermore, data from all participant interactions and the modified chatbot were downloaded from GPTtrainer, counting up to 2546 singular responses.

Starting with the Qualtrics data, we wanted to consolidate the data from the two experiments to gain better insight and merge them into one data frame. After merging the data from the two experiments, we removed 183 responses with missing values under Ideas. Another 39 responses were removed due to irrelevant answers or incomplete questionnaires. We evaluated the removal of irrelevant ideas as justified as these ideas were likely not generated with the intention of experiment but rather to win the awards. Further, we removed 18 ideas from the Control Group as they were made using AI, despite the participants being explicitly told not to use any external aids, including AI tools. Although some ideas were incomplete and arguably not launchable, we included incomplete responses as they are a normal consequence of time pressure (De Paola et al., 2016). Each response was given an ID, randomized, and sent a number between 21 - and 52 ideas in individual Excel files for the judges to rate. After that, ratings were merged and connected to the idea ID.

A crucial part of answering our research question was how AI could foster creativity through co-creation. Therefore, we downloaded all interactions from the GPT chatbot. To control for responses not using the predetermined chatbot, we matched each idea from the experiment with an interaction with the AI assistant. We decided to remove all ideas not generated using the GPTtrainer, as they lacked critical data from the interaction. Including these responses could also create an external addition of variability. We removed 14 observations due to suspicion that another AI was used. After cleaning the questionnaire and controlling the interactions, the final dataset comprised 396 responses.

As all questions in the questionnaire had predetermined options, we saw no need for a discretionary assessment of data cleaning. Furthermore, the only outliers in the dataset were in the time used on the questionnaire as a whole and not in generating the ideas. As this time has no implications on the data, we chose to keep the outliers in the analysis. Finally, all data sets were combined in an Excel file for further analyses in R, assigning all variables to their respective idea ID.

3.3.4 Sample Characteristics

The selection consists of different demographic groups. To examine whether the different groups had the same characteristics, we chose to examine the three groups in more detail. An overview of all the participants' distribution across different participant groups can be found in Table 2.

	Control	**Treatment 1**	**Treatment 2**	**Total**
	Control	Treatment1	Treatment2	Total
N	122	139	135	396
%	30.8%	35.1%	34.1%	100%

Table 2 - Distribution between participant groups

We needed to examine the demographics across the groups further, as we needed to check whether the participants were equally distributed across the control and treatment groups. Table 3 shows that more male participants than females were present across all groups, especially in the Control Group and Treatment Group 2. However, the experiment distribution is somewhat similar to the gender distribution of the Norwegian School of Economics, implying that there was no relatively more male partitioning. Furthermore, the table reveals that the majority of participants are aged 24 - 26 and that there are almost as many participants aged 21-23, and the smallest group is those aged 27 or older. When comparing the three treatments and the number of years studying, fewer participants have studied for three years and more for four years for Treatment Group 2. The finding makes sense as we promoted the last experiment in two lectures for the first year on masters, but no lectures for the third year Bachelor students.

The table additionally demonstrates a clear overweight of students attending the study field of Business/Economics, with 328 participants studying Business/Economics. As we are unaware of potential snowball sampling effects and the exact extent of distribution on social media platforms, we do not know how many potential participants were given the opportunity to participate. At NHH, however, 2955 students received the initial email and the reminder, resulting in 324 NHH students represented as participants in the final dataset. Furthermore, there is an overweight of NHH students in Treatment Group 2, a feature we will discuss further in limitations. Lastly, most participants did not have previous business-oriented work experience.

Variable		Control	Treatment 1	Treatment 2	Total
Gender	*Female*	29%	44%	28%	**140**
	Male	32%	31%	37%	**255**
	Another	0%	0%	100%	**1**
Age	*17 - 20*	19%	48%	33%	**54**
	21 - 23	33%	36%	31%	**155**
	24 - 26	34%	29%	36%	**163**
	27 and above	21%	38%	42%	**24**
Years Studying	*1*	23%	35%	42%	**62**
	2	27%	48%	24%	**62**
	3	42%	42%	17%	**53**
	4	30%	26%	44%	**110**
	5	35%	34%	31%	**74**
	6 + above	29%	31%	40%	**35**
Line Of Study	*Business / Economics*	28%	33%	39%	**328**
	Another	44%	47%	9%	**68**
Work Experience	*No Experience*	32%	37%	31%	**247**
	Relevant Experience	29%	32%	40%	**149**

Table 3 - Demographics of the treatment groups

3.4 Measures

3.4.1 Dependent Variables

We measure creativity using Consensual Assessment Technique (CAT), letting business and investment experts evaluate the ideas (Amabile, 1982; Said-Metwaly et al., 2017). Therefore, we contacted a wide range of people: venture capitalists, entrepreneurs, crowd founders, business leaders, and others with relevant experience, inviting them to act as judges for the experiment. A total of 28 people accepted the role. For the complete list of judges and their relevant experience, see Appendix 8.2.

To avoid priming the judges and potentially affecting the scoring, the judges were not given information about the treatment given to the participants or the purpose of our study (Meyer & Schvaneveldt, 1971). Moreover, they did not know what we wanted to measure (creativity) and were told to evaluate the ideas based on the given scoring criteria. The judges were told, however, that they were to score business ideas generated by students. Further, they were given all the information provided to the participants, except the treatment. Thus, the judges' expectations were modified as they knew the students had only 15 minutes to generate their ideas.

All ideas were evaluated by three judges. Each judge was given 52 ideas to evaluate, though minor adjustments to the idea samples were made to meet all judges' schedules. To reduce potential biases from particularly harsh or soft judges, we implemented a rolling distribution system ensuring that no judge evaluated the same sample of ideas as another judge. Furthermore, the idea order was randomized before distribution, ensuring the judges received a mixed sample of ideas generated with no AI, with AI, and with prompted AI. Due to miscommunication with some judges, nine ideas were evaluated by two judges.

The ideas were evaluated based on two main criteria: creativity and profitability. Both main criteria consisted of two additional criteria, evaluating each idea across four dimensions. All four dimensions were scored on a scale from 1 to 10, with 1 representing the lowest and 10 the highest score possible. For the exact scoring criteria, see Appendix 8.2. The criteria were based on a previous experiment by Boussioux et al. (2023), where they tested the quality of circular business opportunities. Though the judges were told to use the scoring criteria actively, they were also encouraged to score subjectively and make their own considerations, increasing the subjective element of the assessment in line with CAT (Amabile, 1982).

Based on our discussion in the literature review, creativity was divided into novelty and usefulness, as these two concepts are widely mentioned in various creativity definitions (e.g., Batey, 2012; Runco & Jae-ger, 2012; Zeng et al., 2009). To better measure the aspects of creativity, novelty and usefulness were given separate scores. Novel was defined as "to what degree the idea is original and unique concerning whether or not it is similar to an existing idea in any shape or form," while useful was defined as "to what degree the idea solves a real problem, how feasible the idea is, and how comprehendible the solution is." Furthermore, the judges were not told to give an overall creativity score but to evaluate novelty and usefulness separately. The overall creativity score for an idea consists of the idea's average score of novelty and usefulness, which we calculate.

As our primary purpose was to investigate creativity, we wanted to separate the creativity and profitability measures of the business ideas, thus isolating the creativity aspect. Doing so also provides a more nuanced view of idea quality, as profitability is important when generating business ideas (Girotra et al., 2023; Boussioux et al., 2023). Potential substantial differences in the profitability of the ideas warrant further analyses in addition to our analysis of the creativity scores.

As with creativity, profitability was divided into two criteria, and the judges did not give a total profitability score but two separate scores for each dimension. These dimensions were "profit potential" and whether the idea "reached the intended target group." Boussioux et al. (2023) evaluated the financial value of the ideas as one dimension only; however, we opted for a more nuanced approach, splitting the financial value into two dimensions.

The total profitability score was calculated by 0.75 * ProfitPotential + 0,25 * TargetGroup. We placed more emphasis on ProfitPotential as we believed it would be relatively easy to create a product that reached the intended target group and that ProfitPotential is more crucial for an idea's financial value. Our beliefs were underpinned by the idea scores, as TargetGroup received significantly higher scores than the other dimensions. As the relevant market (the average Norwegian student) is particular and small, we considered reaching the average Norwegian student key for ideas to be profitable and therefore wanted to include the variable.

Novel	Useful	ProfitPotensial	TargetGroup	Creativity	Profit	TotalScore
				0.5 Novel	0.75 ProfitPotensial	0.5 Creativity
				0.5 Useful	0.25 TargetGroup	0.5 Profit

Table 4 - Dependent variables

The scoring provided us with measures for novelty, usefulness, profitability, and reaching the intended target group. The two formers represent the ideas' creativity value and, thus, the creativity variable. We also measure profitability, comprising the two latter dimensions, thus representing a profitability variable. All four dimensions and the total scores for creativity and profitability were used as dependent variables in our analyses. We also calculated the "total score" for each idea by taking the average of the creativity and profitability scores, resulting in seven dependent variables used in our analyses. All dependent variables are exhibited in Table 4.

3.4.2 Independent Variables

The independent variable in our experiment is AI access. Subsequently, the only difference between the treatment and Control Groups is access to the AI model ChatGPT-4. In the original experiment distributed through the first survey, AI was measured in binary: the participants were either given access to AI or not. AI access was simply access to the model and no further information or exordium. There was no prompting for how to best interact with the tool or other instructions that aided the participants in generating their ideas using AI.

The second treatment group, however, was given more input on how to best utilize the AI tool for creative purposes. The input included an encouragement to use and interact with the AI and five guidelines for doing so best. Thus, we measure AI access with no guidelines, and we measure AI access with encouragement and guidelines for usage.

3.4.3 Moderation Variables

Domain-specific competence within business was measured in two ways: cognitive competence and functional competence. Both competence measures were applied as moderator variables in separate models. Cognitive competence was measured using students' GPAs for the business degree they are currently pursuing. For business students at the bachelor's level, students' competence was therefore dependent on their GPA at the bachelor's level, while master students' competence was dependent on their GPA at the master's level. Subsequently, students who do not currently pursue a business and or/economics degree were measured to have zero cognitive business competence.

Functional competence was measured using students' time spent studying business/economics. Participants who had studied for four, five, or more years were considered highly competent, while first-year students were considered less so. Non-business/economics students were measured to have zero functional business competence.

GPA was chosen as the cognitive competence measurement for three reasons. First, as discussed in the literature review, Bradley et al. (2007) suggest that GPA and HOCS are correlated. Second, GPA represents a combination of students' intelligence and technical skills (Brown & Campion, 1994; Ickes et al., 1990), contributing to business competence (Tucker & McCarthy, 2001). Third, measuring GPA is easy and provides an adequate measure for our moderator variable. GPA was measured in the survey through self-reporting.

Time spent studying was chosen as the functional competence measurement for two reasons. First, as discussed in the literature review, time spent in a line of studying may increase functional competence (Le Deist & Winterton, 2005; Drejer, 2000). Second, as with GPA, studying time is easy to measure and suitable for our purposes. Time spent studying was also measured through self-reporting in the survey.

3.4.4 Control Variables

To test for differences between demographic groups, we collected participant data concerning age and gender. Furthermore, we measured the time spent answering the experiment and the number of interactions the participants had with the chatbot for both treatment groups. We also collected information concerning participants’ AI experience, technology adoption rate, and openness to new experiences as a predictor for creative behavior as we wanted to investigate whether these variables could explain some variation in the dependent variables. Moreover, we created dummy variables for work experience within business, for the responses only having one interaction with the chatbot, and for direct copies of the task description. All the abovementioned variables were applied as control variables in some of our regression models. For an overview of all variables, their names, and meanings, see Appendix 8.3.

3.5 Data Analysis

3.5.1 Descriptive Statistics

To get an overview of the data, we created tables for descriptive statistics and boxplots, enabling us to compare the variables' means, standard variations, and percentiles. This gave us a good foundation for deciding on further analyses and provided insights into the distribution of the ideas across all seven dependent variables. Furthermore, the plots and tables provide a practical presentation of our initial findings (Saunders et al., 2019). Descriptive statistics were created using R-studio and formatted in Excel.

3.5.2 T-Tests

To test Hypotheses 1 and 2, we tested for significant differences between the means of the three participant groups by conducting several two-sample t-tests in R-studio. As we are interested in whether the treatments result in more creative ideas than the Control Group, one-tailed t-tests are appropriate (Pillemer, 1991). Thus, we can detect significant statistical differences in a positive direction, determining whether Treatment Group 1 was more creative than the Control Group and whether Treatment Group 2 was more creative than Treatment Group 1. To determine significance, we applied a confidence interval of +/- 1.96, terming a significant difference should the p-value be 0.05 or lower, in line with the standard for t-testing (Saunders et al., 2019). We assumed equal variance when conducting the tests after visually inspecting the scatterplots in Appendix 8.5.

We conducted t-tests to investigate differences between the control and two treatment groups across all seven dependent variables, giving us three pairings: the Control Group < Treatment Group 1, Treatment Group 1 < Treatment Group 2, and Control Group < Treatment Group 2. The t-test results were the basis for further analyses using regression models.

Further, as we aim to reveal the potential practical implications of applying AI as a tool, we tested for time spent generating an idea, considering time restrictions may be essential when working. Therefore, we conducted three additional one-tailed t-tests on time spent generating an idea: the Control Group > Treatment Group 1, Treatment Group 1 < Treatment Group 2, and the Control Group > Treatment Group 2.

Moreover, for Treatment Groups 1 and 2, we conducted t-tests for “user interactions” (how many interactions a specific participant had with the chatbot) and for “one response” (the number of participants who had only one interaction with the chatbot). When conducting the tests, we hypothesized that participants in Treatment Group 1 had fewer interactions on average and more one-interaction responses.

3.5.3 Regression Analyses

To further nuance the results achieved in the t-tests, we used R-studio to conduct several OLS regression analyses, allowing us to investigate the coefficient of determination for our variables and gain insight into what may cause variation in creativity (Saunders et al., 2019). OLS was chosen as it is efficient and widely accepted for fitting linear statistical models (Hayes & Chai, 2007). Though we built several models, all took the standard OLS form of the following:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 CV_i$$

Y_i represents one of the seven dependent variables, while β_0 represents the constant. The variable used as the constant varied across the models but was always one of NoAI, NoPromptAI, or PromptAI. The two X variables represent the participant groups not used as the constant, while they represent the various control variables applied in the models.

Before conducting our analyses, we must ensure that all assumptions for multiple linear regression are satisfactory. Subsequently, we tested for linearity, normality, heteroscedasticity, autocorrelation, and multicollinearity (Berry, 1993). Normality was tested by creating Q-Q plots and visually inspecting whether the output was satisfactory. Linearity and heteroscedasticity were tested by plotting the residuals vs. fitted values in scatterplots and inspecting whether the residuals were randomly scattered around the horizontal axis and whether the residuals resembled constant variance. We tested autocorrelation through a Durbin-Watson test, rejecting the null hypotheses of autocorrelations greater than zero as none of the p-values were close to the limit of 0,05 (Saunders et al., 2019). Finally, multicollinearity was tested by calculating the collinearity statistics for tolerance and its inverse; variance inflation factor (VIF), with the former needing to exceed 0,1 and the latter being below 10 to be accepted (Saunders et al., 2019). To meet the collinearity requirements, our final models omitted FirstClick, TimeOfIdeaSubmit, and NumberOfClicks as they highly correlated with the more significant TimeOfLastEdit. Furthermore, TotalYearsStudying and YearsOnEarlierStudy were also omitted as they correlated with one of our moderators, LengthOfCurrentStudy.

Though assumptions were tested for all OLS models we built, we have included the respective assumption tests of two models in Appendix 8.5: one for regressions on the whole population with control variables and one for regressions on the AI population, also with control variables. All assumption tests were run in R-studio.

As we have seven dependent variables and several control variables, there were many potential analyses we could run in addition to testing our hypotheses concerning creativity only. To broaden our understanding of the most significant relationships, we used the t-tests and the descriptive analysis results to narrow our focus to the significant and close-to-significant relationships. However, we first ran regression models testing for all control variables independently across all seven dependent variables, investigating the control variables' effects and significances. Variables that only affected the treatment groups, such as ExperienceAI and UserResponses, were omitted from the models as including them could create an imbalance in the data and complicate the comparison of the Control Group with the treatment groups.

After testing for control variables, we built our final model and conducted six analyses to investigate how the Control Group and the treatment groups affected the dependent variables, always omitting one group to avoid perfect multicollinearity. Subsequently, some models applied NoAI and NoPromptAI as control variables, some applied NoAI and PromptAI as control variables, and some applied NoPromptAI and PromptAI as control variables. All variations of the models were run with and without additional control variables and applied to the population of 396 responses.

Moreover, we conducted additional regression analyses on the treatment groups to investigate how AI affects creativity. When doing so, we applied the AI-specific variables as control variables, testing the effects of the number of interactions with the chatbot, directly copying the chatbot's output, self-reported AI experience, and other variables. The procedure followed the same structure as when we tested for the whole population; first, we tested all control variables' isolated effects and then built models with and without control variables.

3.5.4 Moderation Analyses

To test Hypothesis 3, we conducted moderation analyses in R-studio, investigating how cognitive and functional competence moderated AI's effect on creativity and other dependent variables. The moderation analyses were run as OLS regression models, adding a moderator variable consisting of AI * Competence, in line with the appropriate procedure for testing for moderation effects (Baron & Kenny, 1986). The regression models took the following general form:

$$Y_i = \beta_0 + \beta_1 AI + \beta_2 Competence_i + \beta_3 AI * Competence + \beta_4 CV_i$$

As we are interested in whether competence moderates the effect of AI regardless of how AI is used, we combined Treatment Groups 1 and 2 into a single independent variable called AI. After that, we tested for the potential moderation effects of cognitive and functional competence, measured through GPA and years of studying in the current school, respectively.

Models were built with and without the control variables applied in our regular regressions, represented by C.

Moderation effects were tested by adding an interaction term consisting of the product of AI and the respective Competence interactions while controlling for the AI and Competence variables, thus improving the interpretation of the regression coefficients (Fairchild & MacKinnon, 2009). Potential moderation effects are signaled by the moderation interaction's effect and significance, revealing whether AI affects creativity equally across different competence levels (Baron & Kenny, 1986).

3.5.5 Analysis of the Best Ideas

As ideation often focuses on the best ideas rather than the average idea quality, we conducted additional analysis by thoroughly examining the judges' ratings to ascertain the most effective treatment in high-quality idea generation. To find the most effective treatment, we first had to find who generated the best-rated ideas. We, therefore, organized the ratings across all seven dependent variables in descending order, facilitating the identification of the top ideas for each variable. To illustrate what rating the best idea for each variable within each participant group achieved, we made a line chart. Moreover, we wanted to study whether the ratings of the best ideas indicated the same trends as the top 10 best ideas. Subsequently, we counted the number of ideas from each treatment within the top ten. The process gave us insights into trends, creating a more nuanced analysis. The small sample size made conducting causal hypothesis testing pointless, though our analysis indicates what participant group generated the best ideas.

3.5.6 GPTrater

On November 6th, OpenAI introduced a new feature allowing users to create a custom version of ChatGPT-4. As previous studies found GPT-2 to give reliable predictions of human creativity ratings (Luchini et al., 2023), we hypothesized that GPT-4, being more advanced, would offer even more reliable predictions. Furthermore, the launch offered a method of objectively rating all ideas and evaluating whether our results were consistent in idea evaluation when compared to GPT-raters. Hence, we made two AI judges using the custom version of ChatGPT-4. We only examined the two dependent variables for Creativity, Novel and Useful, as these measurements are documented to be well-imitated when using AI judges (Johnson et al., 2022; Luchini et al., 2023). Moreover, we only wanted to test for creativity, and the custom GPT-4 gave us an efficient tool to isolate the creativity of an idea without being influenced by other factors. For evaluation, the ideas had to be manually copy-pasted into the chatbot. As this process was time-consuming, we only evaluated novelty and usefulness, as creativity is our primary dependent variable.

The first AI rater, hereafter GPTrater-1, was prompted to imitate a human judge. To mimic the judges most accurately, we used 25 random ideas and their associated ratings to train the chatbot. GPTrater-1 was further instructed to use these ratings to evaluate how novel and useful other business ideas were. After that, we fine-tuned the chatbot through iteration by training it on random ideas from the experiment and providing guidance on what aspects should stipulate high and low ratings. The prompt and ideas used to train the chatbot are found in Appendix 8.6. For the second chatbot, hereafter GPTrater-2, we applied a zero-shot approach by prompting the models with the guidelines to give ratings between 1 and 10 but not providing any examples.

When the chatbots were ready, we made them evaluate 100 random ideas. The ideas were the same for both chatbots. We then calculated the means for the chatbots' scores and compared them to each other and the human-given scores. Based on the results, we used one-tailed t-tests to test for differences between the chatbots and between each chatbot and the scores given by the human judges (HJ).

3.6 Evaluation of Validity and Reliability

To ensure our findings are credible and we can draw meaningful conclusions from our analyses, we need to meet the criteria for validity and reliability (Saunders et al., 2019). Subsequently, we outline the criteria and evaluate our study accordingly, thus assessing the quality of our research. We begin by examining the experiment's internal validity before discussing our findings' external validity and reliability.

3.6.1 Internal Validity

Internal validity concerns whether the experimental design captures the desired phenomenon through planned manipulations of the treatment groups (Saunders et al., 2019). Isolating the effects of our independent variables is, therefore, key to achieving high internal validity. To conclude causal effects, high internal validity is required to ensure our measurements are accurate, resulting in credible and trustworthy findings.

We applied random assignment when assigning participants to the control or treatment groups. Consequently, there should be minimal pre-existing differences between the three groups, and the risk of biased results is thus reduced (Saunders et al., 2019). Furthermore, the sample consists solely of Norwegian students, with the vast majority attending the Norwegian School of Economics. As such, our sampling technique can be considered homogeneous, further reducing participant differences. In extension, potential explanations for variations in creativity, other than the treatments, are reduced, strengthening the internal validity of our experiment.

Though we intended to use students only as participants, we could not ensure that all participants were students. Non-students participating may have achieved deviating results as they have different experiences than students and less knowledge concerning what products students demand. Such effects would complicate the isolation of AI as a predictor variable, consequently making it harder to measure the intended effects and reducing internal validity (Saunders et al., 2019). However, non-student participation is unlikely because the experiment was distributed solely to students and in student-only channels. Therefore, it is fair to assume our sample is homogeneous.

With randomized groups from a homogeneous sample, the only difference between the control and treatment groups should be the two AI treatments. Thus, AI access and AI access with guidelines and encouragement, representing planned interventions, are the only explanations for variations in the dependent variable, strengthening the internal validity (Saunders et al., 2019). As such, we isolated the treatment effects.

However, we did not conduct a lab experiment where we could control all variables (Saunders et al., 2019). Though an online experiment was easy to run and gave us a satisfactory sample size, we sacrificed control of potential external variables affecting the participants. We could not enforce participants abiding by the rules and guidelines, and subsequently, we could not ensure they acted as we intended. For example, some participants may have collaborated, and some may have used external aids. Moreover, as the survey was open for a longer duration of time, some students might have talked about the experiment. This could have led to some participants beginning to think of business ideas before the experiment, thus representing a threat to internal validity. However, as the idea's scores did not determine the rewards, we assume students spent little time thinking about ideas beforehand, reducing the potential threat. Furthermore, we had no control over the participants' surroundings. Some may have been fully committed and participated in a quiet room, while others may have participated during lectures, only paying half attention to the experimental task. Subsequently, different circumstances may affect the participants' output. Thus, the lack of control over external effects represents a substantial threat to the internal validity of our findings.

Further, misunderstanding, lazy, and non-attentive participants threaten the experiment's internal validity. Subsequently, we took measures to ensure that participants understood the instructions and exerted sufficient effort. For each page in the survey, participants were told to pay close attention and confirm their understanding of the task. Moreover, as participants were rewarded for participation and not performance, we believed some people would put minimal effort just to enter the reward pool. To counter this, participants had to spend at least two minutes on the experimental task before continuing, and the response had to be a minimum of 10 characters. Though we believe our measures worked as intended, the lack of supervision means we cannot know if some participants misunderstood the assignment or hurried through the survey.

3.6.2 External Validity

External validity can be conceptualized as to what extent our results are generalizable and can be applied to a population beyond our specific research setting (Saunders et al., 2019). Therefore, for our findings to be transferable and bear meaning to other populations, we must ensure sufficient levels of external validity.

To strengthen the external validity, we aimed to maximize the sample size. Subsequently, we carefully planned our recruitment strategy: obtaining desirable rewards and conducting thorough recruitment by sending emails and reminders, entering popular lectures to promote, and encouraging our social network to distribute the survey further to their respective friends and associates. Our efforts resulted in 651 responses and a final sample size of 396, sufficient to achieve significant results for our purposes (Green, 1991). Thus, the number of participants strengthens the external validity.

The external validity is reduced, however, as the experiment applies a homogenous sample using students only as participants, with most participants attending the Norwegian School of Economics. As discussed, a homogenous sample is desirable when isolating effects; however, we cannot know if our findings apply to a broader population or simply to our specific sample. Thus, applying a homogenous sample reduces the generalizability of potential findings, weakening our study's external validity (Saunders et al., 2019).

Furthermore, opting for a non-probability sample negatively affects the study's representativeness (Saunders et al., 2019). As participating is voluntary, students may decide not to participate should they consider themselves uncreative or unmotivated to devote enough time to the survey. However, the average time spent on the survey was only 9,6 minutes, indicating low necessary effort to participate. Further, when deciding whether to participate, we consider it likely that the chances of winning rewards outweigh the participants' intrinsic motivation to be creative. If so, uncreative students may not abstain from participating, reducing the negative effects of applying a non-probability sample, as uncreative people participate regardless.

3.6.3 Reliability

To ensure high reliability and credible findings, the experiment must produce consistent results that are easy to replicate, and the survey must be comprehensible for the participants to avoid misunderstandings (Saunders et al., 2019). Typical threats to reliability include participant and researcher biases or errors. Consequently, we took several measures to reduce potential participant errors and biases, as well as errors and biases on our part.

To ensure the experimental task and the questionnaire were understandable, we did rigorous testing using friends as test participants. The original experimental task was based on previous research conducted by Girotra et al. (2023); however, we somewhat altered the formulations through iteration processes with the test participants. The test participants also provided feedback on the questionnaire, ensuring all questions were straightforward with understandable scales. For the relevant measures, for example, self-reported creativity, we applied empirically tested and validated scales, like the Big Five, to ensure proper measurement of the constructs (Schrauf & Navarro, 2005). Thus, the risk of participants misunderstanding questions and/or the experimental task was reduced.

Nevertheless, as discussed in the evaluation of internal validity, we did not supervise the participants. Consequently, the circumstances surrounding each participant may have varied and will be hard to replicate, reducing the continuity of our results (Saunders et al., 2019). Further, no supervision makes detecting and accounting for participant errors hard. Subsequently, potential replication attempts of our study may struggle to produce similar results, reducing the reliability of our findings.

We do not know if some people participated more than once. To ensure anonymity, we did not collect IP addresses, emails, or any other constructs connecting responses to participants. Though participants were told they could only submit a singular response, some may have participated more than once to increase their chances of winning a reward. A second response would cause biased results as that idea would not be affected by the treatment in the intended way and would produce unreliable results (Saunders et al., 2019). Thus, as we could not enforce purely singular responses, multi-response participants threaten the credibility and reliability of our study.

When collecting and preparing the data, we were careful to avoid researcher bias and errors by constantly discussing our choices and sanity-checking with friends to get an outside perspective. As discussed in Preparation of the Data, we removed incomplete and unserious responses, as those would damage the reliability of our results. When doing so, we carefully evaluated why a response was deemed incomplete or unserious, double-checking whether time spent on the survey or other factors underpinned our suspicions. For example, an idea that did not meet any of the requirements stated in the participation prompt was theorized to be “unserious.” If the time spent generating the idea was at the minimum (two minutes), we believed the respondent simply wanted to enter the reward pool and hurried his or her response without regard to performance. Though we systematically approach the data, errors may have been made, and either fewer or more responses could potentially have been removed. However, all removed responses were discussed and evaluated several times to ensure the correct decision was made to strengthen the reliability of our results.

3.7 Research Ethics

Research Ethics concerns the standards for how researchers behave regarding the rights of all subjects to their research (Saunders et al., 2019). Ethical concerns are evident for the whole duration of a research process and are greater when research involves human participants, irrespective of whether the participation occurs online or is conducted person-to-person.

To ensure adequate standards, we carefully read and followed the guidelines from FEK and strived to abide by them throughout the research process (forskningsetikk.no, 2019). For our study, ethical implications are mainly connected to anonymity, confidentiality, and the handling of personal data of those affiliated with our experiment.

Our data concerns human participants and was obtained through internet-mediated access (Saunders et al., 2019). The survey was developed based on Datatilsynet's and FEK's guidelines, thus being in line with what both institutions consider sufficient informed consent regarding participation (Datatilsynet, 2015; forskningsetikk.no, 2019). When distributing the survey, we were careful to press that participation was voluntary. Further, we took measures to keep participants anonymous and remove all doubts participants may have had regarding their anonymity. Anonymity was achieved by not collecting IP addresses and not asking any identity-revealing questions.

As entering a reward pool represented an incentive for participation, we needed to decide on four winners, thus revealing their identities. After a response was registered, the participant would be redirected to a Google Forms and given an option to enter the reward pool by providing his or her email address. The participants were instructed that entering the reward pool would reveal their identity, though we clarified that their identity could not be linked to their response. Thus, most participants were revealed to us, though we did not gain access to sensitive personal data. Ideally, all participants should remain completely anonymous; however, as confidentiality and informed consent were upheld, we consider the Google Forms an adequate compromise to maximize sample size while simultaneously considering research ethics.

Moreover, we applied a set of human judges to evaluate the ideas generated by the participants and thus gained elite person access (Saunders et al., 2019). When recruited, the judges were made aware that accepting the role was voluntary and that they did so to aid two students in writing their master's thesis. Further, all judges were asked whether we could present their names and relevant experience in line with confidentiality and informed consent principles.

4. Results

4.1 Descriptive Statistics

4.1.1 Descriptive Statistics for whole population

To gain insight into the data, we incorporate descriptive statistics, including features from the questionnaire and GPTtrainer, divided into the control and the treatment groups. Reviewing Table 5, it is evident that all the average ratings for all our dependent variables are at or below the middle score of 5. This suggests that the evaluators have been conservative in their assessments. Notably, Novelty stands out, with Treatment Group 2 showing a slightly higher average score of 0.2 compared to the others, a difference we plan to investigate further using t-tests. ProfitPotential received the lowest scores across the board and was, on average, almost two points under TargetGroup. In contrast, the measure for TargetGroup received the highest average score, at 5, coupled with higher 25th and 75th percentiles. The standard deviations, ranging from 1.05 to 1.41 across the groups, suggest that the elevated average score does not result from a wide dispersion of ratings or multiple high-rating clusters.

The Control Group's mean scores do not differ much from the treatment groups; however, they received, on average, the highest score for Useful, Profit, and ProfitPotential. This is an exciting feature that we will investigate with t-tests. For the Control Group, the statistical deviations of TotalScore, Creativity, Profit, ProfitPotensial, and TargetGroup were also higher than average, while Novel and Useful were on average. The deviations imply that ratings for ideas generated without co-piloting were more spread out, in line with theory.

		N	Mean	Std. Dev	Pctl25	Pctl75
Novel	*Treatment 1*	139	3.76	1.24	2.83	4.5
	Treatment 2	135	4.04	1.21	3	4.67
	Control	122	3.69	1.3	3	4.33
Useful	*Treatment 1*	139	4.29	1.26	3.5	5.33
	Treatment 2	135	4.37	1.41	3.33	5.33
	Control	122	4.4	1.38	3.5	5.33
ProfitPotensial	*Treatment 1*	139	3.03	1.23	2	3.59
	Treatment 2	135	3.07	1.05	2.33	3.67
	Control	122	3.13	1.25	2	4
TargetGroup	*Treatment 1*	139	4.93	1.26	4.09	6
	Treatment 2	135	4.93	1.25	4	5.67
	Control	122	5.02	1.41	4	6
Creativity	*Treatment 1*	139	4.03	1.09	3.25	4.79
	Treatment 2	135	4.2	1.16	3.33	5
	Control	122	4.05	1.14	3.5	4.67
Profit	*Treatment 1*	139	3.51	1.12	2.71	4.08
	Treatment 2	135	3.54	1.01	2.83	4.11
	Control	122	3.6	1.19	2.64	4.41
TotalScore	*Treatment 1*	139	3.77	1.04	3.06	4.29
	Treatment 2	135	3.87	1.01	3.15	4.5
	Control	122	3.82	1.08	3.25	4.41

Table 5 - Descriptive analysis of the dependent variables

Figure 4 illustrates that the scores for novel and creativity are more spread out for the Control Group. Moreover, more extreme values are in the Control Group, represented as outlier dots. However, their lower and higher whisker is more comprised, suggesting that the scores are more similar than the treatment groups.

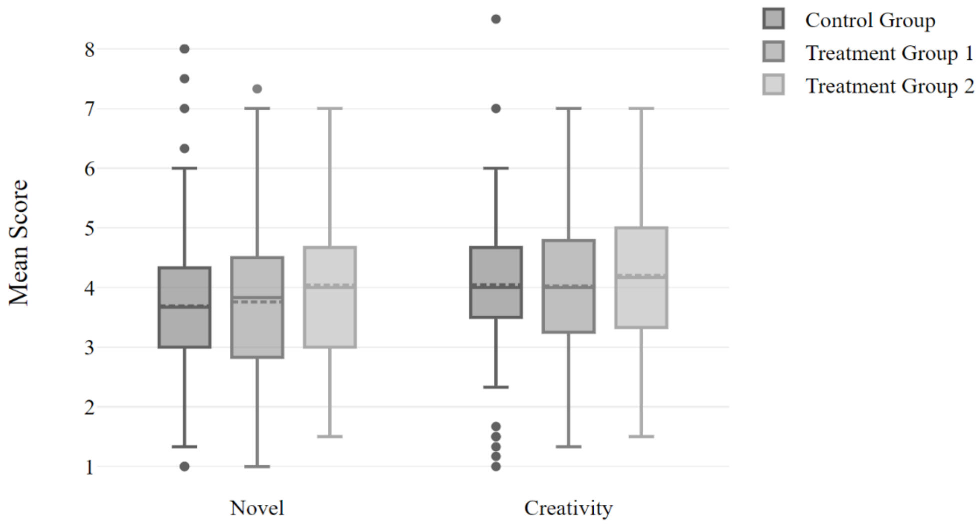


Figure 4 - Spread for Novel and Creativity

When only focusing on the treatment groups, there is a clear pattern. Treatment Group 2 received better or equal ratings across all dependent variables. This suggests that Treatment Group 2 generated more creative ideas with more economic potential than the ideas generated by Treatment Group 1.

The Control Group used, on average, almost seven minutes to generate their idea, as shown in Table 6. The high average time is likely due to the participants needing to describe their ideas without tools. Interestingly, Treatment Group 1 generated almost as well-rated ideas in almost two and a half minutes shorter time. Besides, the guidelines stipulated that Treatment Group 2 used longer to generate ideas. Furthermore, the Control Group had more clicks, suggesting they iterated their idea more times in the questionnaire. It also indicated that the treatment groups iterated their ideas more using the chatbot.

		N	Mean	Std. Dev	Pctl25	Pctl75
TimeOfLastEdit	*Treatment 1*	139	4.58	3.26	2.02	3.8
	Treatment 2	135	6.53	4.51	2.14	5.83
	Control	122	6.81	4.22	3.4	5.89
NumberOfClicks	*Treatment 1*	139	11.17	11.72	4	13.5
	Treatment 2	135	10.44	9.45	4	13
	Control	122	20.46	21.26	7	27.75

Table 6 - Time in minutes and number of interactions

4.1.2 Descriptive for AI Treatment

Examining AI-related attributes for the treatment groups, the data indicates that most students started using ChatGPT last year. Furthermore, most students had prior experience in AI and were open to using new technology. The means are also similar, indicating no significant difference between the treatment groups.

		N	Mean	Std. Dev	Pctl25	Pctl75
FirstTimeUsingAI	*Treatment 1*	139	4.77	1.24	4	6
	Treatment 2	135	4.75	1.18	4	6
ExperienceAI	*Treatment 1*	139	3.03	0.88	2	4
	Treatment 2	135	3.14	0.86	3	4
NewTechnology	*Treatment 1*	139	3.7	1	3	4
	Treatment 2	135	3.89	0.93	4	4.5

Table 7 - Descriptive statistics concerning AI and Technology

Moving on, Treatment Group 2 interacted almost twice as much as Treatment Group 1, indicating that they used more interactions to generate their idea. This implies that the guidelines altered the behavior of Treatment Group 2. Furthermore, around 43% of Treatment Group 1 and 22% of Treatment Group 2 sent one message to the chatbot before delivering their idea, indicating that they went with the first idea that the chatbot generated. Concerning the first interaction with the chatbot, participants in both treatment groups directly copied the task description into the interaction 50% of the time, indicating that they wanted the chatbot to generate the idea and not iterate an existing idea.

		N	Mean	Std. Dev	Pctl25	Pctl75
UserResponses	*Treatment 1*	139	2.46	1.96	1	3
	Treatment 2	135	4.17	2.91	2	6
One_Response_D	*Treatment 1*	139	0.43	0.5	0	1
	Treatment 2	135	0.22	0.42	0	0
DirectCopy_D	*Treatment 1*	139	0.47	0.5	0	1
	Treatment 2	135	0.48	0.5	0	1

Table 8 - Descriptive statistics for chatbot use

4.2 T-tests

To test for Hypothesis 1 and Hypothesis 2, we conducted two-sample one-tailed t-tests. Hypotheses 1 and 2 read as follows:

H1: Human-AI co-creation positively affects creativity.

H2: Human-AI co-creation produces more creative ideas when the user efficiently prompts and interacts with the AI.

First, H1 was tested by examining whether the Control Group < Treatment Group 1 for Creativity. As there were no significant differences between the two sample means when testing for Creativity at a 5% level, we can reject H1.

Second, we tested H2 by conducting another t-test: Treatment Group 1 < Treatment Group 2 for Creativity. The test was significant at a 10% level, indicating that appropriate prompting and interaction with the chatbot might produce more creative ideas; however, given no significant differences at a 5% level, we must reject H2.

Further t-testing revealed that Treatment Group 2 produced more novel ideas than the Control Group and Treatment Group 1; both tests yielded significant results at a 5% level. Though our hypotheses stipulated effects on creativity and not novelty, novelty is a crucial aspect of creativity, and the significant differences are thus findings that we will elaborate on in our regression analysis. We also tested for all the other dependent variables but found no significant differences. The participant groups' respective means and standard errors for novelty and creativity can be viewed in Figure 5.

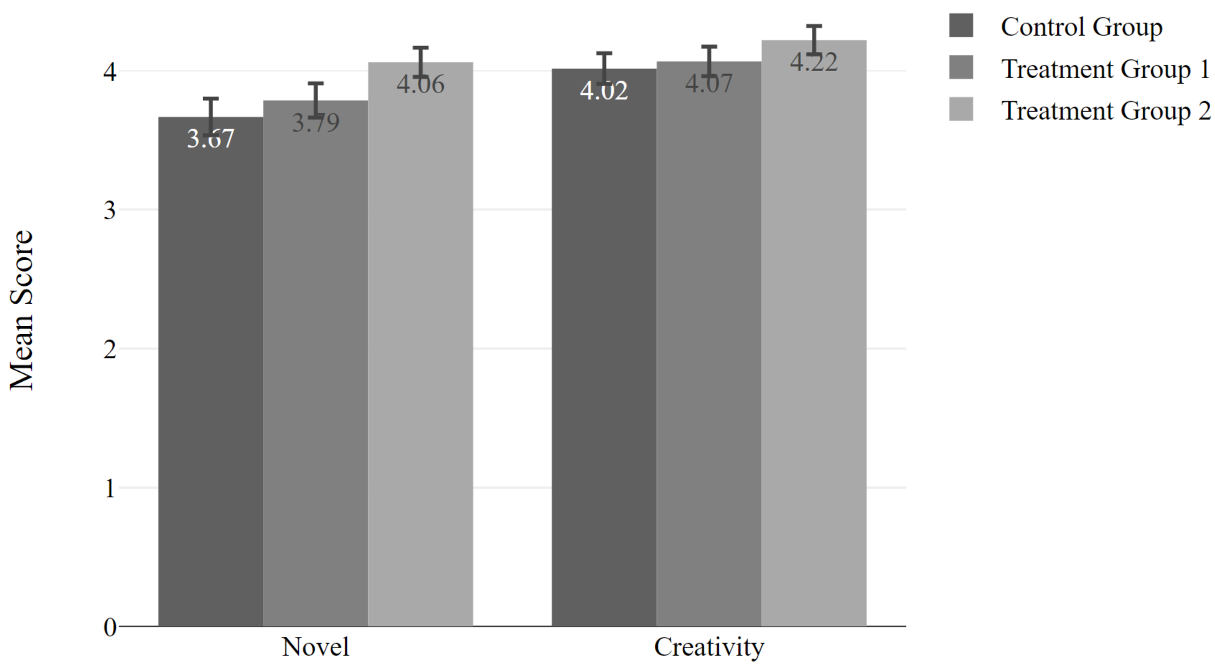


Figure 5 - Mean scores and standard errors for Novel and Creativity

All in all, our initial analyses made through descriptive statistics and t-testing indicate no significant differences in any of the seven dependent variables except for the novelty of the ideas produced by Treatment Group 2. Thus, ideas generated by the Control Group are of equal novelty and creativity as ideas generated with unprompted AI users and of equal creativity as ideas generated with prompted AI users.

Moving on to the t-tests concerning control variables, there were significant differences between Treatment Group 1 and the Control Group and Treatment Group 2 in time spent generating the ideas. As showcased in Figure 6 (TimeOfLastEdit is presented in minutes), participants in Treatment Group 1 were much faster, and all the t-tests yielded p-values of $p < 0.001$. Further, there was no significant difference between time spent between Treatment Group 2 and the Control Group, though we observed that Treatment Group 2 spent slightly less time on average.

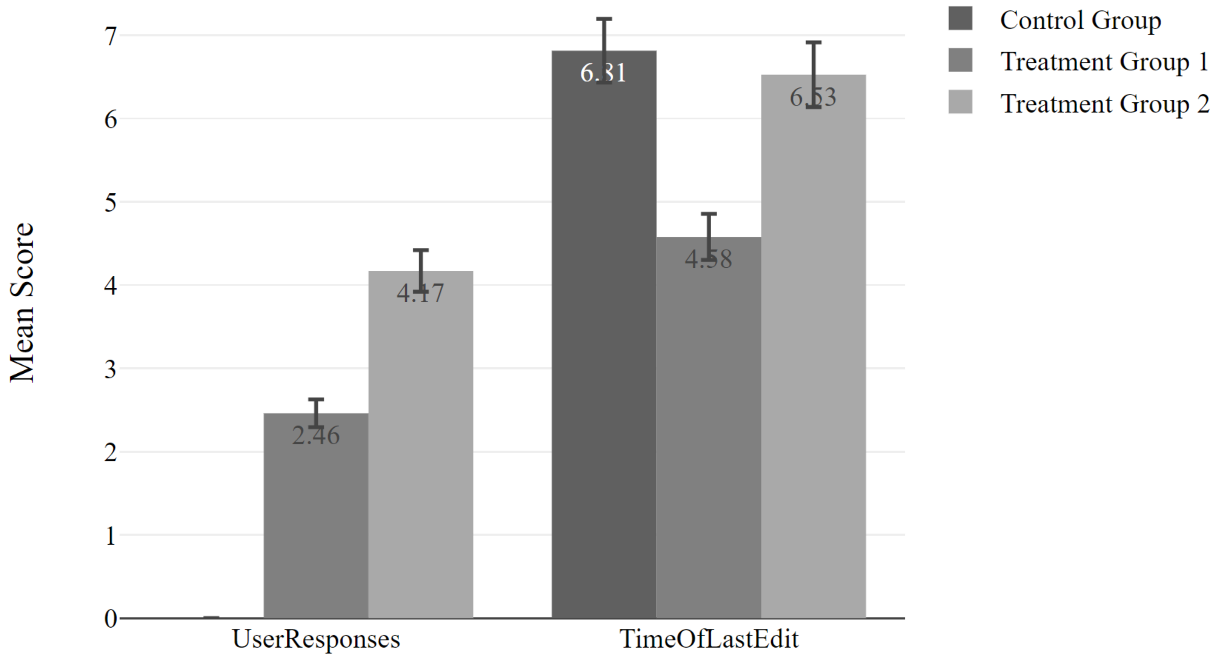


Figure 6 - Mean scores and standard errors for time and interactions

Further t-testing for Treatment Groups 1 and 2 revealed that unprompted AI users had significantly ($p < 0.001$) fewer interactions with the chatbot than the prompted AI users, as exhibited in Figure 4. Moreover, there were significantly ($p < 0.001$) more participants in Treatment Group 1 who only had one interaction with the chatbot, indicating that the prompting of Treatment Group 2 had the intended effect. Thus, Treatment Group 2 prompts and interacts with the chatbot more. For all t-test results, see Appendix 8.4.

4.3 Regression Analyses

The t-tests stipulated further investigation of the differences in Novelty between the participant groups. Hence, we investigated the control variables' coefficients in isolation, using the whole population of 396, applying Novel as the dependent variable. The results are shown in Table 9. Apart from the constant, the only significant results at a 5% level were the effects of PromptAI and TimeOfLastEdit, both being significant in isolation and when control variables were applied. PromptAI having a significant effect was no surprise given the t-test results and seemed to increase the novelty score by approximately 0.35. TimeOfLastEdit, however, achieved an estimate slope coefficient of near zero, indicating that, though significant, the variable had little to no effect on novelty scores.

Further, in line with the t-tests, we observe that the first treatment yielded no significant results, though it did have a mildly positive effect. The results indicate that using AI can create more novel ideas; however, the effect may depend on how AI is used.

Interestingly, TotalYearsStudying, which represents one of our measures for competence, was significant at a 10% level when all control variables were included. The effect was mildly positive, indicating that people with high functional competence generate more novel ideas. However, our measure for cognitive competence, GPA, has a negative estimated slope coefficient, though the effect is small and insignificant. The results of our moderation analyses are presented later in the section under Moderation Analyses.

	Dependent variable:							
	Novel							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
PromptAI_D	0.347^{**}	0.351^{**}	0.360^{**}	0.367^{**}	0.358^{**}	0.345^{**}	0.343^{**}	0.370^{**}
	(0.156)	(0.156)	(0.155)	(0.156)	(0.156)	(0.155)	(0.156)	(0.157)
NoPromptAI_D	0.075	0.067	0.140	0.084	0.060	0.074	0.070	0.146
	(0.156)	(0.155)	(0.158)	(0.155)	(0.155)	(0.154)	(0.155)	(0.160)
Gender	0.055							0.053
	(0.133)							(0.139)
AgeInterval		-0.015						-0.101
		(0.079)						(0.092)
TimeOfLastEdit			0.001^{**}					0.001^{**}
			(0.0003)					(0.0003)
LengthOfCurrentStudy				0.065				0.088^{*}
				(0.044)				(0.053)
GradeAverage					-0.011			-0.061
					(0.126)			(0.129)
SelfReportedCreativity						0.057		0.065
						(0.041)		(0.043)
WorkExperience							0.097	0.037
							(0.130)	(0.138)
Constant	3.653^{***}	3.726^{***}	3.475^{***}	3.498^{***}	3.735^{***}	3.268^{***}	3.656^{***}	3.175^{***}
	(0.144)	(0.225)	(0.155)	(0.173)	(0.509)	(0.322)	(0.122)	(0.623)
Observations	395	396	396	396	394	396	396	393
R^2	0.015	0.015	0.025	0.020	0.016	0.020	0.016	0.040
Adjusted R^2	0.007	0.007	0.017	0.012	0.008	0.012	0.009	0.017
Residual Std. Error	1.249 (df = 391)	1.247 (df = 392)	1.241 (df = 392)	1.244 (df = 392)	1.248 (df = 390)	1.244 (df = 392)	1.246 (df = 392)	1.243 (df = 383)
F Statistic	1.966 (df = 3; 391)	1.958 (df = 3; 392)	3.337^{**} (df = 3; 392)	2.661^{**} (df = 3; 392)	2.074 (df = 3; 390)	2.607^{*} (df = 3; 392)	2.135^{*} (df = 3; 392)	1.764^{*} (df = 9; 383)
Note:								$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 9- Control variables for whole population

To examine the effects more closely, we built four models to investigate the participant groups' effect on novelty compared to each other, as shown in Table 10. In line with the results from the t-tests and previous regressions, we observe that PromptAI has a significant positive effect, shown in Models 1 and 2, when the Control Group is omitted from the models. The effect was evident without and with control variables, Models 3 and 4 omit NoPromptAI from the model to investigate the effects of PromptAI and NoAI without and with control variables. As confirmed by t-testing, the results show that NoAI has a mildly negative and insignificant effect and that PromptAI leads to more novel ideas than NoPromptAI, though only significant at a 10% level. Further, the effect becomes insignificant when applying control variables, though the estimated slope coefficient is still positive, indicating that Treatment Group 2 indeed produced more novel ideas. Simultaneously, TimeOfLastEdit is still significant at a 5% level in Model 4. Taken together with the t-test showing that Treatment Group 2 spent significantly more time generating ideas than Treatment Group 1, the results indicate that how AI is used may affect the ideas' novelty scores.

	Dependent variable:			
	Novel			
	(1)	(2)	(3)	(4)
PromptAI_D	0.351^{**}	0.370^{**}	0.282^{*}	0.224
	(0.156)	(0.157)	(0.151)	(0.157)
NoPromptAI_D	0.069	0.146		
	(0.155)	(0.160)		
NoAI_D			-0.069	-0.146
			(0.155)	(0.160)
TimeOfLastEdit		0.001^{**}		0.001^{**}
		(0.0003)		(0.0003)
Gender		0.053		0.053
		(0.139)		(0.139)
AgeInterval		-0.101		-0.101
		(0.092)		(0.092)
LengthOfCurrentStudy		0.088^{*}		0.088^{*}
		(0.053)		(0.053)
GradeAverage		-0.061		-0.061
		(0.129)		(0.129)
SelfReportedCreativity		0.065		0.065
		(0.043)		(0.043)
WorkExperience		0.037		0.037
		(0.138)		(0.138)
Constant	3.690^{***}	3.175^{***}	3.759^{***}	3.321^{***}
	(0.113)	(0.623)	(0.106)	(0.612)
Observations	396	393	396	393
R^2	0.015	0.040	0.015	0.040
Adjusted R^2	0.010	0.017	0.010	0.017
Residual Std. Error	1.246 (df = 393)	1.243 (df = 383)	1.246 (df = 393)	1.243 (df = 383)
F Statistic	2.927^{*} (df = 2; 393)	1.764^{*} (df = 9; 383)	2.927^{*} (df = 2; 393)	1.764^{*} (df = 9; 383)
Note:				$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table 10 - Regression analysis whole population

To investigate if the nature of how AI is used affects novelty, we conducted regression analyses on the AI population only, applying AI-specific variables as control variables. The results are shown in Table 11. We observe that PromptAI is significantly more novel than NoPromptAI in all models but two. The first is when UserResponses is applied as a control variable, and the other is when all control variables are included. Considering PromptAI is significant for all other control variables in isolation, the results indicate that some of the variations caused by PromptAI may be explained by UserResponses, that is, how many interactions the participants had with the chatbot.

Further, we observe that the dummy variable for participants only interacting with the chatbot once (One_Response) has a negative B coefficient when tested in isolation. Considering the positive effect of UserResponses, this comes as no surprise, though One_Response becomes positive when all control variables are included in Model 7. Model 7, however, achieves the lowest adjusted R-squared; subsequently, we should not put too much emphasis on the model.

Examining the user's AI attributes reveals that when the user first used AI, the user's AI experience and technology adoption rate all affect novelty positively, though insignificantly. This effect is strongest for FirstTimeUsingAI - having a positive effect in isolation and in combination with the other control variables. Moreover, we observe that the model applying FirstTimeUsingAI only achieves the highest R-squared.

	Dependent variable:						
	Novel						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PromptAI_D	0.284^{*}	0.279^{*}	0.279^{*}	0.221	0.272^{*}	0.281^{*}	0.223
	(0.147)	(0.148)	(0.149)	(0.156)	(0.152)	(0.148)	(0.158)
FirstTimeUsingAI	0.082						0.100
	(0.061)						(0.073)
ExperienceAI		0.024					-0.038
		(0.086)					(0.104)
NewTechnology			0.012				-0.022
			(0.077)				(0.086)
UserResponses				0.035			0.048
				(0.030)			(0.038)
One_Response_D					-0.045		0.067
					(0.161)		(0.202)
DirectCopy_D						0.061	0.078
						(0.148)	(0.154)
Constant	3.366^{***}	3.687^{***}	3.713^{***}	3.672^{***}	3.778^{***}	3.730^{***}	3.296^{***}
	(0.310)	(0.279)	(0.303)	(0.127)	(0.125)	(0.125)	(0.405)
Observations	274	274	274	274	274	274	274
R^2	0.020	0.014	0.013	0.018	0.014	0.014	0.028
Adjusted R^2	0.013	0.006	0.006	0.011	0.006	0.007	0.002
Residual Std. Error	1.219 (df = 271)	1.223 (df = 271)	1.223 (df = 271)	1.220 (df = 271)	1.223 (df = 271)	1.223 (df = 271)	1.226 (df = 266)
F Statistic	2.734^{*} (df = 2; 271)	1.855 (df = 2; 271)	1.829 (df = 2; 271)	2.524^{*} (df = 2; 271)	1.855 (df = 2; 271)	1.903 (df = 2; 271)	1.084 (df = 7; 266)
Note:							*p<0.1; **p<0.05; ***p<0.01

Table 11 - Control variables for AI population

4.4 Moderation Analyses

Hypothesis 3 was tested using moderation analysis, investigating competence's potential moderating effects on the AI-creativity relationship. Hypothesis 3 reads as follows:

H3: AI affects creativity differently, depending on the individual's domain-specific competence, regardless of how AI is used.

In addition to analyzing AI and creativity, we considered it appropriate to present the potential effects for Novelty, as this was the only dependent variable where the t-test yielded significant results. The first four models apply Novel as the dependent variable, and the last four apply Creativity as the dependent variable. All models are shown in Table 12. We also tested the moderation effects for the other dependent variables and found no significance for any of them, except TargetGroup moderated by LengthOfCurrentStudy. The moderation effect was significant at a 5% level; however, as we focus on creativity, we do not further elaborate on this.

	Dependent variable:							
	Novel				Creativity			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI_D	0.665*	0.607	0.211	-0.065	0.568*	0.530	-0.513	-0.730
	(0.365)	(0.368)	(1.538)	(1.538)	(0.327)	(0.325)	(1.373)	(1.360)
LengthOfCurrentStudy	0.160*	0.150		0.084	0.159*	0.130		0.052
	(0.096)	(0.099)		(0.057)	(0.086)	(0.088)		(0.050)
TimeStudying_Moderation	-0.127	-0.091			-0.140	-0.108		
	(0.111)	(0.112)			(0.099)	(0.099)		
GradeAverage		-0.170	-0.117	-0.263		-0.216	-0.304	-0.423
		(0.149)	(0.350)	(0.354)		(0.132)	(0.313)	(0.313)
Grade_Moderation			0.014	0.101			0.161	0.236
			(0.383)	(0.384)			(0.342)	(0.339)
Gender		0.236		0.241		0.180		0.191
		(0.155)		(0.156)		(0.137)		(0.138)
AgeInterval		-0.039		-0.046		0.060		0.050
		(0.102)		(0.102)		(0.090)		(0.090)
TimeOfLastEdit		0.001**		0.001**		0.001***		0.001***
		(0.0003)		(0.0003)		(0.0002)		(0.0002)
SelfReportedCreativity		0.052		0.055		0.041		0.044
		(0.047)		(0.047)		(0.041)		(0.041)
WorkExperience		0.023		0.026		-0.002		0.002
		(0.146)		(0.146)		(0.129)		(0.129)
Constant	3.180***	3.194***	4.137***	3.749**	3.529***	3.588***	5.231***	4.636***
	(0.321)	(0.752)	(1.408)	(1.467)	(0.287)	(0.665)	(1.257)	(1.297)
Observations	328	326	327	326	328	326	327	326
R^2	0.019	0.047	0.011	0.045	0.014	0.061	0.009	0.059
Adjusted R^2	0.010	0.020	0.002	0.018	0.005	0.034	0.0001	0.032
Residual Std. Error	1.241 (df = 324)	1.238 (df = 316)	1.247 (df = 323)	1.239 (df = 316)	1.110 (df = 324)	1.094 (df = 316)	1.113 (df = 323)	1.095 (df = 316)
F Statistic	2.072 (df = 3; 324)	1.724* (df = 9; 316)	1.237 (df = 3; 323)	1.655* (df = 9; 316)	1.527 (df = 3; 324)	2.284** (df = 9; 316)	1.008 (df = 3; 323)	2.202** (df = 9; 316)
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 12 - Moderation analysis

As neither TimeStudying_Moderation nor Grade_Moderation significantly affected Creativity, we found no support for H3. Effects were also insignificant when tested for novelty. Interestingly, TimeStudying_Moderation had a negative estimate slope coefficient, while Grade_Moderation had a positive one, indicating that cognitive competence might increase AI's effects, while functional competence might reduce it. However, the results are not significant, and the models without control variables achieve low adjusted R-squared. Therefore, our findings indicate that both competencies have no moderation effects on how AI affects creativity and novelty and that AI affects creativity equally, independent of domain-specific competence.

4.5 Additional Analyses

4.5.1 Best Ideas

To investigate which of the three participant groups generated the most creative ideas, we examined the best idea in each group more closely. Considering that our primary analysis indicated that Treatment Group 2 produced the most novel ideas, with no significant differences between the other dependent variables, we might expect similar results when investigating the best ideas. However, the best-rated idea was generated in the Control Group across all categories, including Novel and Creativity. For a comparison of the best idea from each participant group, see Figure 7.

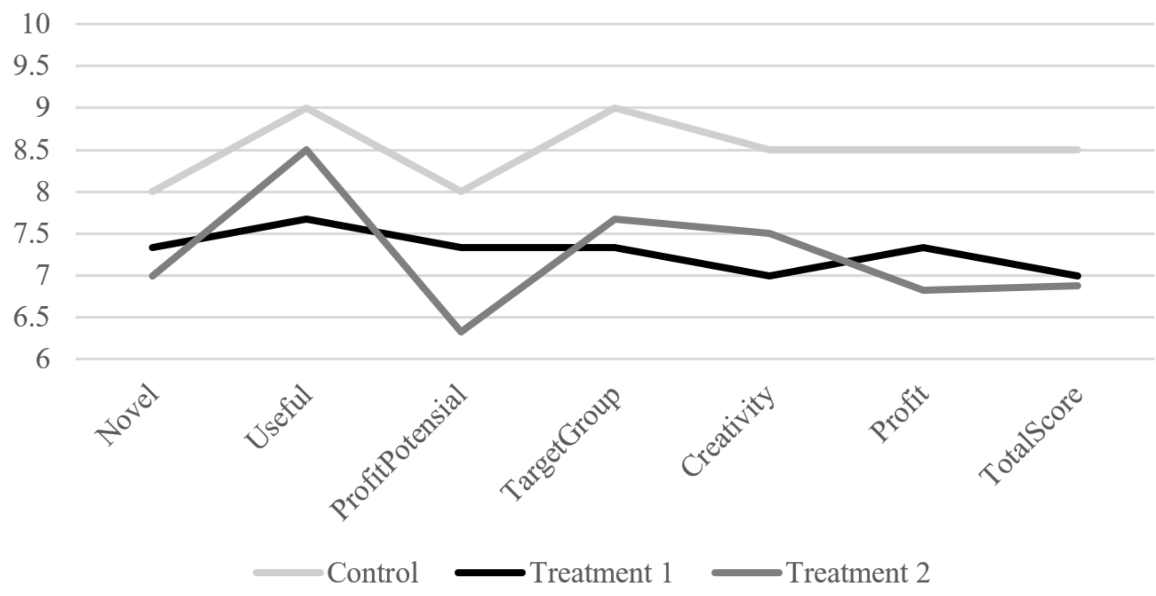


Figure 7 - Best ideas across populations

Furthermore, when examining the three best-rated ideas for all dependent variables, two of the three best-rated ideas among all variables are from the Control Group. Moreover, the Control Group generates the three most novel ideas. The findings suggest that the best ideas are generated when highly creative individuals are free to explore without limits and not co-creating by AI.

However, when examining the top ten best-rated ideas across the variables, most are from Treatment Group 2. This trend may indicate that while the Control Group excels in generating exceptionally high-rated ideas, Treatment Group 2 fosters the development of a more significant number of quality ideas through co-creation. This could suggest that the guidelines applied in Treatment Group 2, although not producing the highest creativity peaks as seen in the Control Group, are effective in consistently nurturing above-average creative outputs.

4.5.2 GPTratings

Examining the means for scores given by GPTrater-1, GPTrater-2, and the human judges (HJ) made us hypothesize that HJ were the most harsh, followed by GPTrater-1. Subsequently, we tested for HJ < GPTrater-1 and GPTrater-1 < GPTrater-2 in terms of given scores for Novel, Useful, and Creativity. The means and standard errors are exhibited in Table 13. The t-tests showed no significant difference between GPTrater-1 and HJ for Novel, but the chatbot gave significantly higher scores for usefulness ($p < 0.001$). Further, to no surprise when inspecting the means, GPTrater-2 gave significantly higher scores ($p < 0.001$) for both dependent variables when compared to GPTrater-1.

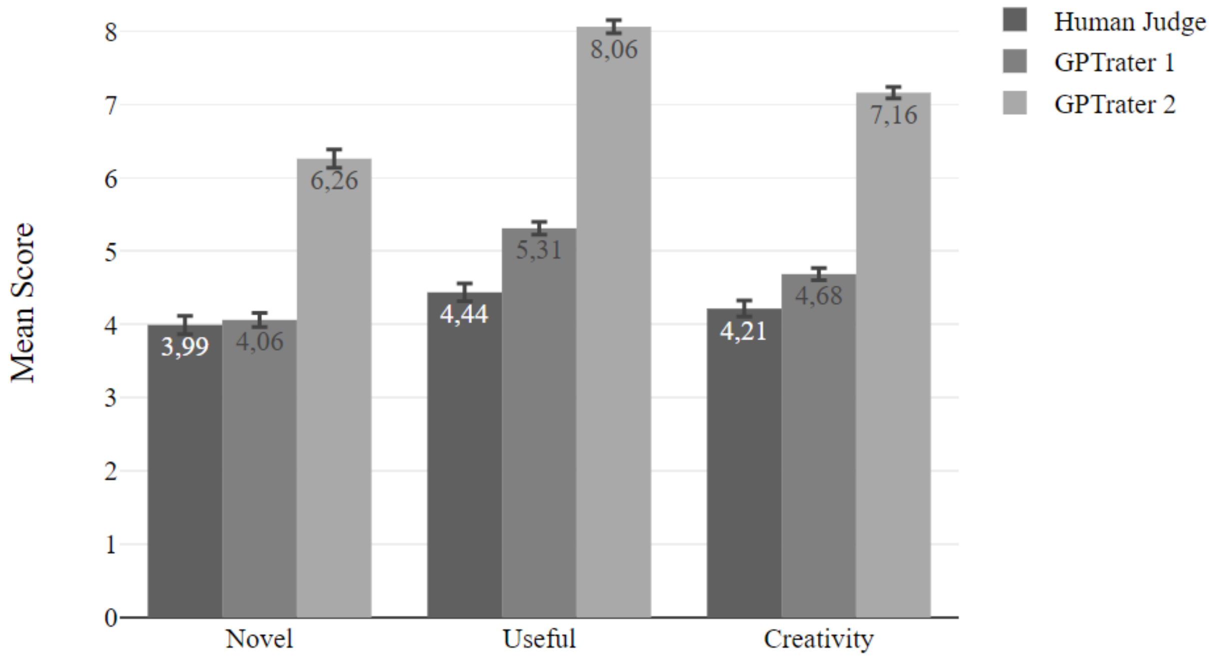


Figure 8 - Mean scores given with different evaluation methods

5. Discussion

5.1 Theoretical Contributions, Implications, and Future Research

5.1.1 Main Findings

This master thesis applied a deductive approach, an explanatory study design, and an experimental method to answer the following research question:

“Does human-AI co-creation positively affect creativity, and does the level of competence moderate the relationship?”

We contribute to the current literature (e.g. Haase & Hanel, 2023; Boussioux et al., 2023; Girotra et al.; 2023; Wan et al., 2023), offering a more nuanced understanding of the AI-creativity relationship. Contradicting previous research on AI-creativity in isolating, we found no significant difference between the overall creativity for non-AI and the non-prompted human-AI co-creating participants. At first glance, it seems co-creating with AI does not necessarily stipulate more creative outcomes.

However, when interacted with correctly, AI increases the novelty of creative outputs and should subsequently be considered as a tool for creative knowledge workers within business. While we found no significant overall differences in general creativity, the prompted AI users generated significantly ($p < 0.05$) more novel ideas than non-AI participants. Furthermore, the groups had an almost identical score for usefulness, meaning the increased novelty scores did not negatively affect the viability of the ideas. Our results resonate with the findings of Wan et al. (2023), who suggest that co-creating with AI to process and recombine a vast array of information in unique ways can lead to novel outputs (Wan et al., 2023; Doshi & Hauser, 2023). This suggests a shift in AI's role from a supportive background tool to a more proactive partner when creating novel outputs in general. AI should not replace human novelty; instead, it might introduce a new dimension to the dynamic, where AI's data-driven insights complement human intuition and experience. As such, the interaction between AI and humans is essential for an equal partnership rather than a one-sided dependency.

Taking our results together with the findings of Wan et al. (2023) and Dell’Acua et al. (2023), we emphasize the importance of iterative interaction in the co-creative process. Merely co-piloting with AI is insufficient; it is the active, directive involvement of humans that unleashes the novel potential of AI. This finding also resonates with Anantrasirichai & Bull's (2022) view of AI as an augmentative tool, implying that it is not AI alone but how we use it that enhances novelty. Therefore, contrary to the general assumption of AI as a universal enhancer, our study advocates for a more nuanced understanding. It is not about whether AI is used but how it is used that determines the novel output.

Consequently, businesses should not blindly adopt AI as a creative tool to be used carelessly, highlighted by our finding that Treatment Group 1 did not produce more novel results than the Control Group and significantly ($p < 0.05$) less novel ideas than Treatment Group 2. Our regression analysis indicates that the number of interactions might explain the differences in novelty between the co-piloting groups, as the prompted individuals had significantly ($p < 0.001$) more interactions. The number of user responses was also the only variable to make the difference between prompted and unprompted groups insignificant, implying it explains some of the difference for novelty. Further investigation is needed, however, to understand better how to apply the tool to harness its creative powers. Subsequently, future research should experiment with various approaches to gain insight into how to make AI produce novel outputs through co-creation. This research should include experimentation with temperatures, as temperature is a determent for AI's creative output. We applied the standard temperature; however, higher temperatures produce more random outputs and can thus be more creative.

Though our results suggest effective human-AI co-creation significantly boosts novelty, these findings challenge the traditional view of AI in creative processes; AI enhances ideas' usefulness, but humans are more novel (Boussioux, 2023; Kirkpatrick, 2023; Haase & Hanel, 2023). Our contradicting findings lead us to consider whether our experiment produces results that differ in novelty and usefulness compared to previous research because previous studies tested AI in isolation and not through co-creation. Alternatively, our results could be influenced by the rapid advancements in AI technology or the specific demographics of our participants, being mainly skilled students. The latter factor calls for replicating our experiment and applying a different sample. Further, to better understand the implications of the rapid development of AI technologies, future studies should compare three groups: participants without AI, AI in isolation, and participants co-creating with AI. Thus, one can examine whether the differences in novelty stem from the nature of co-creation, the rapidly evolving AI technology, or participant characteristics.

Furthermore, our results indicate that AI can work as a productivity enhancer for creative work, given that both AI groups generated ideas in a shorter time than the control group. AI’s effectiveness came as no surprise, considering the models excel at creating content quickly (Wolfram, 2023). For businesses that require a high volume of creative content, for example, content marketing, AI's time efficiency means that more content can be produced within a given timeframe, potentially scaling the creative output without additional human labor. As an extension, AI might allow creative professionals to allocate more time to more complex aspects of their work, like refining concepts or fine-tuning ideas to better align with strategic goals. Thus, co-piloting with AI could work as a tool for clearing routine creative tasks to make way for more complex problem-solving.

Further, we highlight that Treatment Group 1 was significantly faster than Treatment Group 2, indicating that fewer interactions may be advantageous when time represents a substantial limitation. Though thorough iteration produces more novel outputs, and we found that time spent had a small positive, yet significant effect on novelty, fewer interactions may produce satisfactory results within a shorter time frame. Subsequently, fewer interactions can be favorable when time represents a limitation, and the output only needs to meet certain requirements. However, we did not manipulate time pressure and cannot conclude on time pressure’s effects. To gain further insights, future research should investigate how to best apply AI as a creative tool dependent on available time.

Moving on, our findings stipulate that all knowledge workers can use AI to produce more novel outputs, regardless of their competence within their respective domains. We found no moderating effects for domain-specific competence concerning novelty, usefulness, or creativity, suggesting that highly competent knowledge workers should apply AI to the same degree as less competent individuals. Our findings, however, contrast the results of Noy & Zhang (2023) and Dell’Aqua et al. (2023), who found that competence had a leveling effect, enhancing the quality most for less competent workers. A reason for our deviating results could be our participants being a homogenous group of highly competent people - having an excellent GPA from high school to attend the Norwegian School of Economics (NHH). Hence, the cognitive competence should be somewhat the same for the whole group.

Furthermore, as our moderator sample consists of business students and our experiment is not complicated from a business perspective, our experiment could have been more directed at knowledge acquired throughout the study. As such, the experiment might not have been optimal to find competence's moderating effect. Therefore, it might be of interest for future research to facilitate an experiment with participants with a normal distribution of cognitive competence and test for more specific domain competence.

Furthermore, the assessment of traditional competence metrics might have overlooked the growing role of AI literacy in creative collaboration (Ilomäki, 2011; Wang & Yuan, 2022). This oversight can be paralleled with the evolution of internet proficiency. Initially, competence in using the internet was a unique skill that set some individuals apart in their ability to access and synthesize information. Over time, however, this skill has transformed from a distinctive advantage to a fundamental necessity in creative work. Similarly, AI literacy is emerging as a possible differentiator for creative professionals. It enables a deeper engagement with AI tools, encouraging novel idea generation by tapping into AI's vast data processing capabilities. Our results underpin this sentiment; when the participants used AI for the first time and their overall experience positively affected the novelty of their ideas, though not significantly.

Nevertheless, as AI technology becomes more integrated into everyday business operations and creative endeavors, the ability to effectively utilize AI will likely evolve from being a differentiator to a hygiene factor. This transition underlines a crucial point: While capturing a snapshot of this transition, our study might have missed these nuances of AI literacy and its growing impact on creative collaboration. Future research should thus investigate what skills are paramount for generating novel and creative outputs in the future and could integrate AI literacy as a competence requirement.

Taken together, we contribute to the research field by showing that human-AI co-creation can produce more novel outputs than humans alone. Our findings have several implications for how knowledge workers should approach future ideation processes. Regardless of competence, this effect is equal for all workers and provides valuable insight for businesses and management.

5.1.2 Additional Findings

Even though our findings indicate that effective human-AI co-creation leads to more novel outcomes on average, previous research emphasizes that organizations seek a few great ideas rather than an abundance of mediocre ones (Girotra et al., 2010). While enhancing average performance is advantageous in routine tasks like landing an airplane, some processes only search for one great output. For example, a pharmaceutical company focuses solely on the most promising drug candidates when deciding its investment strategies. This raises a crucial question: While co-creating with AI can augment the quantity and novelty of ideas on average, does the co-creation process lead to the most creative ideas?

Our findings reveal that the three most novel, the two most useful, and the two most creative ideas came from non-AI participants. The finding is consistent with those of Haase & Hanel (2023), who suggest that the best humans outperform AI regarding the highest creativity scores. The results indicate that the best innovators and creative workers thrive unhindered by AI, as AI may limit the thought process required for the most creative ideas. Therefore, the distinction between AI as an enhancement of overall novelty and a contributor to exceptional creativity in the best ideas is vital.

Though AI can assist in providing diverse perspectives and thus complement creativity, it fails to replace the intuitive and often unpredictable nature of human thought processes that lead to big C creativity. From a management perspective, these findings advocate for a balanced approach to integrating AI into creative processes. Organizations should recognize the value of human creativity and ensure that AI co-creation does not overshadow it. This balance is especially crucial for companies reliant on pioneering innovations, such as pharmaceutical companies, where the over-reliance on AI might lead to a homogenization of ideas and restrain the most innovative thought. That being said, our experiment is not ideal for testing big C creativity, as big C ideas usually require substantial time to develop (Necka et al., 2006). Furthermore, we highlight that we base our discussion on one observation: our sample. Future research must produce similar results before concluding whether the most creative humans consistently outperform human-AI co-creation. It would also be interesting to investigate whether the most creative humans can make use of AI.

However, co-creating with AI could be an effective tool for endeavors in everyday business life that require creative solutions. While the very best ideas came from non-AI users, most ideas within the top ten most creative ideas were generated by Treatment Group 2. The result is reasonable, considering creativity is often displayed as quick thinking and reorganizing acquired knowledge, while the best ideas demand more time and deep domain knowledge (Necka, 2006). Combined with the time efficiency of AI, these results imply that co-creation could be beneficial in contexts where generating a larger volume of feasible ideas is preferred over a few outstanding ones. Thus, in industries where little C is required daily, AI can be a powerful creative tool for knowledge workers as they rely on a constant flow of workable ideas rather than rare, major breakthroughs. Examples are consultants or other problem solvers who face a wide range of challenges that require creative solutions (Sternberg, 1999) or workers in ongoing incremental innovation processes dependent on small breakthroughs.

Moving on to the GPTraters, the scoring patterns between the human judges (HJ), GPTrater-1, and GPTrater-2 reveal key differences in their assessment approaches. GPTrater-1, trained to emulate human judgment, showed a similar level of strictness as HJ in scoring novelty, yet it rated the usefulness significantly higher. The finding suggests that while AI can closely replicate human-like assessment for novel outcomes, it struggles to evaluate the application of ideas in real life. However, GPTrater-2 generally scored higher on both variables, indicating a less constrained evaluation style, in line with its zero-shot learning approach (Johnson et al., 2022).

These differences have implications for improving efficiency and reducing the labor costs of creative assessment processes. The ability of AI raters, particularly GPTrater-2, to evaluate ideas quickly could be leveraged to screen a large volume of ideas swiftly, identifying promising ones for more detailed evaluation by human experts. This approach could substantially reduce the time and labor-intensive efforts typically required in the initial stages of business idea screening (Baer & McKool, 2014; Kaufmann et al., 2007. Furthermore, the method could mitigate some problems concerning subjective biases and constraints for human evaluators. In our study, for example, we spent substantial amounts of time recruiting judges, and the judges spent valuable time evaluating the ideas. Our results indicate that with proper training, the evaluation process could have been partly outsourced to an AI. However, given

AI’s struggles to correctly assess usefulness, human interference is still required at this point in the technological evolution.

Overall, our additional analyses found that the most creative ideas were created by humans only, though we highlighted that these findings must be subject to further investigation. In addition, our findings on the creative assessment capabilities of AI suggest that the tool can be utilized as an initial idea screener, saving knowledge workers substantial time and resources.

5.2 Limitations

Our study is subject to several limitations that should be considered when assessing our results. As we did not provide performance-dependent incentives, participants may not have been stimulated to perform their best. If so, our data set may have been biased as the ideas scores could represent true human-AI co-creation creativity more because of low effort. Should effort indeed be low, idea scores favor AI users, as the chatbot is less dependent on time availability than humans. This may alter our results to indicate that AI is more effective than it is. On one hand, Treatment Group 2 did not spend significantly more or less time than the Control Group, indicating equal effort between the two groups. On the other hand, Treatment Group 1 spent significantly less time than the Control Group, which might indicate lower effort from the unprompted AI users. Lower effort from AI users may represent a potential effect evident in the real world – people get lazier when they have AI access. However, this was not the intended construct of measure, and the lack of incentives potentially caused low effort among participants, thus representing a limitation in our study.

Moreover, as we did not conduct our experiment in a natural setting, we may have created an environment that does not augment purely human creativity, potentially favoring the treatment groups. Human creativity depends on intrinsic motivation (Prabhu et al., 2008; De Jesus et al., 2013; Fischer et al., 2019), and the participants in the Control Group may, therefore, not have performed as well as they would have in a natural setting where they have a source of intrinsic motivation. As with the lack of incentives, a lack of intrinsic motivation may reduce the creativity of the ideas in the Control Group. Ideas co-created with AI will likely not be affected equally, as the AI cannot be motivated. Thus, the unnatural setting of the experiment and the subsequent lack of intrinsic motivation limit our study.

Furthermore, there were fewer students from schools other than the Norwegian School of Economics in Treatment Group 2 compared to the Control Group and Treatment Group 1. Subsequently, the differences between the groups may not have been an effect of effective co-creation but were due to differences between the populations. To test this, we conducted a t-test for novelty, usefulness, and creativity with the Control Group and Treatment Group 1 populations. We split the participants from NHH and other schools and tested for the difference in means between the groups, receiving close to 1 p-values. Furthermore, we tested our hypotheses on NHH students only and got the same results as for the whole population, indicating that it did not largely affect our results.

As discussed in our evaluation of the validity and reliability of the study, we did not supervise the participants and were thus unable to control other factors affecting creativity. For example, participants in the Control Group may have used AI without our knowledge, and participants in all three groups may have used external aids or collaborated with others. Considering we removed 18 ideas (13%) from the Control Group responses, assuming the participants did not follow instructions closely is reasonable. Therefore, there may be bias in our results caused by external effects, and the lack of supervision consequently represents a limitation in our study.

Further, the judges do not have experience or expertise in the relevant market, potentially limiting their ability to give fair scores. To ensure the participants had sufficient knowledge of the market for their product and to make comparing ideas easier, we chose to restrict the ideas to target the average Norwegian student. However, none of the judges had experience in this specific market, and none were recent students. Though the judges qualify for CAT concerning general business and investment experience, their lack of market expertise may have affected their evaluation of ideas targeted at students. Some judges highlighted this limitation, stating that some ideas left them wondering if the product would be desired by students today.

Moreover, some judges highlighted that several ideas from the treatment groups were similar, negatively affecting the score of later ideas. When evaluating the first ideas, the judges were unaware of the similarities and thus initially gave the first ideas higher scores than later ones. Though the judges went back to reevaluate some ideas, similar ideas may still have caused biases in the evaluation and represent a limitation to our study.

Furthermore, limited time and resources reduced our scope for the data analysis. For example, we did not create control variables for the judges or test for significant scoring differences. Accounting for such differences may have altered our results, as some judges may have been harsher than others. Subsequently, a more thorough data analysis would have given a more nuanced view of potential differences between schools and may have affected our results.

Moving on, as we used students as participants, the competence gap may have been too narrow to cause significant moderation effects. The difference in experience between a first- and fifth-year student is only five years, while in practice, the experience gap between new hires and seniors can span several decades. This sentiment gains support, considering none of our moderation measures yielded significant results. Ideally, we should have investigated low and high extremes of business competence; however, recruiting such participants would not have been feasible in our study. Subsequently, complications concerning measuring the moderation effects limit our analysis.

Lastly, first-year students had received few grades at the time of the experiment. Thus, the basis for their GPA variable was limited, potentially causing noise in our data. For example, one of the first courses at NHH is an ethics course; therefore, this may be the only course some first-year students have completed. Subsequently, a student with an A in Ethics, with no other completed courses, will have the highest possible cognitive business competence score. As such, few grades for first-year students limit our GPA variable and subsequent moderation analysis.

6. Conclusion

Our study found that human AI co-creation does not positively affect creativity, regardless of how AI is used. However, closer examination revealed that efficiently prompting and interacting with AI produces more novel ideas, significant at a 5% level. Further, we found that competence had no moderating effects on the AI-creativity relationship, indicating that AI is equally effective across all competence levels. In addition, the two most creative ideas were generated by non-AI users. Taken together, our results contribute to the literature concerning how and when knowledge workers should utilize AI as a creative tool, providing valuable insights for management.

In conclusion, our research on human-AI co-creation in creativity reveals a nuanced message: AI's role in the creative process is to generate a breadth of novel ideas, yet the depth of the most groundbreaking creations remains outside AI's frontier, only achievable for the most creative humans. This interplay suggests a collaborative future where AI is not a substitute for human intelligence but a tool to amplify human creativity and ingenuity. As we look ahead, the opportunity for creative professionals lies in mastering the co-creative process with AI, harnessing its potential not as competing forces but as dynamic partners in creative collaboration. The future of creativity is not human or AI but the uncharted territory of synthesizing minds and machines.

7. References

Acemoglu, D., & Restrepo., P. (2019). Automation and new tasks: How technology. *Journal of Economic Perspectives, 33*(2), 3-30.

Ali, R., Tang, O. Y., Connolly, I. D., Fridley, J. S., Shin, J. H., Sullivan, P. L., & Asaad, W. F. (2022). Performance of ChatGPT, GPT-4, and Google bard on a neurosurgery oral boards preparation question bank. *Neurosurgery, 10*, 1227.

Alu-Mutka. (2011). Mapping digital competence: Towards a conceptual understanding. *Institute for Prospective Technological Studies*, 7-60.

Amabile, T. M. (1982). Social psychology of creativity: A Consensual Assessment Technique. *Journal of Personality and Social Psychology, 43*, 997-1013.

Amabile, T. M. (1983). The social psychology of creativity: A componential conceptualization. *Journal of personality and social psychology, 45*(2), 357.

Amabile, T. M. (1996). Creativity in context. *Boulder, CO: West view Press.* .

Amabile, T. M., & Pratt, M. G. (2016). The dynamic componential model of creativity and innovation in organizations: Making progress, making meaning. *Research in organizational behavior, 36*, 157-183.

Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: a review. *Artificial intelligence review*, 1-68.

Andreassen, T. W., & Ugland, Y. (2023, September 21). *NHH.no*. Retrieved from https://www.nhh.no/en/research-centres/digital-innovation-for-growth/dig-news-and-blogs/2023/how-to-best-spend-one-billion-on-ai/

Ångström, R. C. (2023). Getting AI Implementation Right: Insights from a Global Survey. *California Management Review, 66(1)*, 5-22.

Anton, E. B. (2020). The humans behind Artificial Intelligence–An operationalisation of AI competencies.

Audia, P. G. (2007). Past success and creativity over time: A study of inventors in the hard disk drive industry. *Management Science, 53*, 1-15.

Baer, J., & McKool, S. S. (2014). The gold standard for assessing creativity. *International Journal of Quality Assurance in Engineering and Technology Education, 3*, 81-93.

Baron, R. M. (1986). The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology, 51*(6), 1173.

Barron, F. (1955). The disposition towards originality. *Journal of Abnormal and Social Psychology, 51*, 478-485.

Bass, D. (2023, January 23). *Bloomberg.com*. Retrieved from https://www.bloomberg.com/news/articles/2023-01-23/microsoft-makes-multibillion-dollar-investment-in-openai?embedded-checkout=true

Batey, M. (2012). The measurement of creativity: From definitional consensus to the introduction of a new heuristic framework. *Creativity Research Journal, 24*, 55-65.

Benešová, A. &. (2017). Benešová, A., & Tupa, J. (2017). Requirements for education and qualification of people in Industry 4.0. *Procedia manufacturing, 11*, 2195-2202.

Berry, W. D. (1993). *Understanding regression assumptions.* Sage.

Binsted, K. (1996). Machine Humour: An Implemented Model of Puns. *Ph. D. thesis, University of Edinburgh*.

Boon, J., & van der Klink, M. (2002). Competencies: the triumph of a fuzzy concept. *Academy of HumanResource Development Annual Conference, 1*, 327-334.

Boussioux, L., N Lane, J., Zhang, M., Jacimovic, V., & & Lakhani, K. R. (2023). The Crowdless Future? How Generative AI Is Shaping the Future of Human Crowdsourcing. *The Crowdless Future*.

Boyatzis, R. (2007). Competencies in the 21st century. *Journal of Management Development, 27*(1), 5-12.

Bradley, R. V. (2007). A study on the impact of gpa on perceived improvement of higher-order cognitive skills. *Decision Sciences Journal of Innovative Education, 5(1),* 151-168.

Bradley, S. D. (2012). The relationship between continuing education and perceived competence, professional support, and professional value among clinical psychologists. . *Journal of continuing education in the health professions, 32(1),* 10-21.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems, 33*, 1877-1901.

Bruner, J. S. (1962). The conditions of creativity. *Contemporary Approaches to Creative Thinking, 1958, University of Colorado, CO, US; This paper was presented at the aforementioned symposium,* 1-30.

Brynjolfsson, E. R. (2018). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. *The economics of artificial intelligence: An agenda* , 23 - 57.

Campbell, D. T. (1960). Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological review, 67(6),* 380.

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., & Raffel, C. (2021). Extracting training data from large language models. *USENIX Security Symposium, 21*, 2633-2650.

Carlsson, B., & Eliasson, G. (1994). The Nature and Importance of Economic Competence. *Industrial and Corporate Change, 3*(3), 687-711.

Choi, J. H., & Schwarcz, D. (2023). Ai assistance in legal analysis: An empirical study. *Available at SSRN 4539836.*

Christiaans, H. H. (2002). Creativity as a design criterion. *Creativity Research Jour-nal, 14,* 41-54.

Cohen, H. (1995). The further exploits of AARON, painter. *Stanford Humanities Review, 4*(2), 141-158.

Cope, D. (2005). *Computer Models of Mucial Creativit.* Boston: MIT Press.

Cropley, A. J. (1967). *Creativity.* London: Longmans.

Cropley, D. H., Kaufman, J. C., & Cropley, A. J. (2011). Measuring Creativity for Innovation Management. *Journal of Technology Management & Innovation, 6*(3), 13-30.

Dane, E. (2010). Reconsidering the trade-off between expertise and flexibility: A cognitive entrenchment perspective. *Academy of Management Review*, 579 - 603.

Datatilsynet. (2015). *Anonymisering av peronopplysninger Veileder.* Oslo: Datatilsynet.no.

De Dreu, C. K. (2008). Hedonic tone and activation level in the mood-creativity link: toward a dual pathway to creativity model. *Journal of personality and social psychology, 94(5)*, 739.

De Jesus, S. N., Rus, C. L., Lens, W., & Imaginário, S. (2013). Intrinsic motivation and creativity related to product: A meta-analysis of the studies published between 1990–2010. *Creativity Research Journal, 25*(1), 80-84.

De Paola, M. &. (2016). Who performs better under time pressure? Results from a field experiment. *Journal of Economic Psychology, 53*, 37-53.

Deci, E. L., & Ryan, R. M. (2013). Intrinsic motivation and self-determination in human behavior. *Springer Science & Business Media*.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., & Lakhani, K. R. (2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. *Harvard Business School Technology & Operations Mgt.*

Doshi, A. R., & Hauser, O. (2023). Generative artificial intelligence enhances creativity. *Available at SSRN.*

Drejer, A. (2000). Organisational learning and competence development. *The learning organization, 7(4)*, 206 - 220.

Dreyfus, S. E. (2004). The five-stage model of adult skill acquisition. *Bulletin of science, technology & society, 24(3)*, 177-181.

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130*.

Ennis, R. H. (1985). A logical basis for measuring critical thinking skills. *Educational leadership, 43*(2), 44-48.

Fairchild, A. J., & MacKinnon, D. P. (2009). A General Model for Testing Mediation and Moderation . *Prevention Science*, 87-99.

Felten, E. W., Raj, M., & Seamans, R. (2023). Occupational heterogeneity in exposure to generative ai. *Available at SSRN 4414065*.

Fischer, C. M. (2019). The influence of intrinsic motivation and synergistic extrinsic motivators on creativity and innovation. *Frontiers in psychology, 10*, 137.

Gallardo-Gallardo, E. &. (2021). Talent management for the future of work. In New directions in the future of work. *Emerald Publishing Limited*, 35 - 54.

Gaut, B. (2010). The Philosophy of Creativity. *Philosophy Compass, 5*(12), 1034-1046.

Girotra, K., Meincke, L., Terwiesch, C., & Ulrich, K. T. (2023). Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN 4526071.*

Girotra, K., Terwiesch, C., & Ulrich, K. T. (2010). Idea generation and the quality of the best idea. *Management science, 56*(4), 591-605.

Goldin, C., & Lawrence F., K. (1998). The origins of technology-skill complementarity. *Quarterly Journal of Economics, 3*(4), 693-732.

Goleman, D. (1998). *Working with emotional intelligence.* Bantam.

gpt-trainer.com. (2023). Retrieved from https://gpt-trainer.com/

Green, S. B. (1991). How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research*, 499-510.

Guzdial, M. &. (2019). *An interaction framework for studying co-creative ai.* New York: Cornell University Press.

Guzik, E. E., Byrge, C., & Gilde, C. (2023). The originality of machines: AI takes the Torrance Test. *Journal of Creativity, 33*(3).

Haase, J., & Hanel, P. H. (2023). Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *arXiv preprint arXiv:2303.12003*.

Harris, J. A. (913-929). Measured intelligence, achievement, openness to experience, and creativity. . *Personality and individual differences, 36(4)*, 2004.

Hayes, A. F. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior research methods, 39*, 709-722.

Hingorani, K. S. (1998). Teaching project management through an information technology-based method. *Project management journal, 29(1)*, 10-21.

https://www.forskningsetikk.no. (2019). Retrieved from Generelle forskningsetiske retningslinjer: https://www.forskningsetikk.no/retningslinjer/generelle/

Hu, K. (2023, February 2). *Reuters.com*. Retrieved from https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Ickes, C. &. (1990). Study Skills Differences among High-Risk College Freshmen.

Ilomäki, L. K. (2011). What is digital competence? *Linked Portal*.

Jackson, P. W., & S., M. (1965). The person, the product, and the response: Conceptual problems in the assessment of creativity. *Journal of Personality, 33*, 309-329.

Jo, A. (2023). The promise and peril of generative AI. *Nature (614(1)*, 214-216.

Johnson, P. C. (2022). Digital innovation and the effects of artificial intelligence on firms' research and development–Automation or augmentation, exploration or exploitation? *Technological Forecasting and Social Change, 179*.

Kandlhofer, M. S.-G. (2016). Artificial intelligence and computer science in education: From kindergarten to university. *IEEE frontiers in education conference (FIE)*, 1-9.

Kaufman, J. C. (2016). *Creativity 101.* Springer Publishing Company.

Kaufman, J. C., & Sternberg, R. J. (2007). Resource review:. *Change, 39*, 55-58.

Kaufman, J. C., Lee, J., Baer, J., & Lee, S. (2007). Captions, consistency, creativity, and the Consensual Assessment Technique: New evidence of reliability. *Thinking Skills and Creativity, 2*, 96-106.

King, P., Oxley, M., & King, P. (2001). Individuals and competence. *Housing: Who Decides?*, 9-69.

Kirkpatrick, K. (2023). Can AI Demonstrate Creativity? *Communications of the ACM, 66*(2), 21-23.

Knasel, E., & Meed, J. (1994). Becoming Competent: Effective Learning for Occupational Competence. *Sheffield: Employment Department*.

Kneller, G. F. (1965). *Art and science of creativity.* New York: Holt, Rinehart & Winston.

Le Deist, F. D., & Winterton, J. (2005). What is competence? *Human resource development international, 8*(1), 27-46.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science, 33*(1), 126-148.

Lee, P., Bubeck, S., & Petro, J. (2023). Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine. *New England Journal of Medicine, 388*(13), 1233-1239.

Long, D. &. (2020). What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 . *CHI conference on human factors in computing systems* , 1-16.

Lubart, T. I., & Guignard, J. (2004). The generality-specificity of creativity: A multivariate ap-proach. *In R. J. Sternberg, E. L. Grigorenko, & J. L. Singer (Eds.), Creativity: From po-tential to realization. Washington, DC: American Psychological Association*, 43-56.

Luchini, S. M.-P. (2023). Automatic Scoring of Creative Problem-Solving with Large Language Models: A Comparison of Originality and Quality Ratings.

Mannucci, P. V. (2018). The differential impact of knowledge depth and knowledge breadth on creativity over individual careers. *Academy of Management Journal, 61(5)*, 1741 - 1763.

Mansfield, B. (2004). Competence in transition. *Journal of European Industrial Training, 28*, 296-309.

Manyika, J. C. (2017). . A future that works: AI, automation, employment, and productivity. *McKinsey Global Institute Research, Tech. Rep, 60*, 1-135.

McCallin, A., & Bamford, A. (2007). Interdisciplinary teamwork: is the influence of emotional intelligence fully appreciated? *Journal of Nursing Management, 15*(4), 386-391.

Merrotsy, P. (2013). *A note on big-C creativity and little-c creativity*. Crawley: Creativity Research Journa.

Meyer, D. E. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, 227-234.

Miller, A. I. (2019). The artist in the machine: The world of AI-powered creativity. *MIT Press*.

Mollick, E. (2023, May 9). *Substack.com*. Retrieved from AI is not good software. It is pretty good people: https://www.oneusefulthing.org/p/ai-is-not-good-software-it-is-pretty?utm_source=profile&utm_medium=reader2

Mumford, M. D. (1988). Creativity syndrome: Integration, application, and innovation. *Psychological Bulletin, 103*, 27-43.

Necka, E., Grohman, M., & Slabosz, A. (2006). Creativity studies in Poland. In J. C. Sternberg, *The International Handbook of Creativity* (pp. 270-306). New York: Cambridge University Press.

Nordhaug, O., & Grønhaug, K. (1993). Competences as resources in firms. *The International Journal of Human Resource Management, 5*(1), 89-106.

Norris, N. (1991). The trouble with competence. *Cambridge journal of education, 21*(3), 331-341.

Notar, C. E. (n.d.). Distant learning for the development of higher-level cognitive skills. *Education, 122(4)*, 2002.

Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. *Available at SSRN 4375283*.

openai.com. (2023). Retrieved from https://openai.com/gpt-4

Peng, S., Kalliamvakou, E., Cihon, P., & Demirer, M. (2023). The impact of ai on developer productivity: Evidence from github copilot. *arXiv preprint arXiv:2302.06590*.

Pillemer, D. B. (1991). One-versus two-tailed hypothesis tests in contemporary educational research. *Educational Researcher, 20*(9), 13-17.

Prabhu, V. S. (2008). Creativity and certain personality traits: Understanding the mediating effect of intrinsic motivation. *Creativity Research Journal, 20*(1), 53-66.

Qualtrics.com. (2023). Retrieved from https://www.qualtrics.com/

Reiter-Palmon, R., & Illies, J. J. (2004). Leadership and creativity: Understanding leadership from a creative problem-solving perspective. *The leadership quarterly*, *15*(1), 55-77.

Runco, M. A. (1988). Creativity research: Originality, utility, and integration. *Creativity Research Journal*, *1*(1), 1-7.

Runco, M. A., & Jaeger, G. J. (2012). The Standard Definition of Creativity. *Creativity Research Journal*, *24*(1), 92-96.

Runco, M. A., & Pritzker, S. (1999). Encylcopedia of creativity. *CA: Academic Press*, 71-79.

Said-Metwaly, S., Van den Noorgate, W., & Kyndt, E. (2017). Approaches to Measuring Creativity: A Systematic Literature Review. *Creativity. Theories – Research - Applications*, *4*(2), 238-275.

Saunders, M. N., Lewis, P., & Thornhill, A. (2019). *Research Methods for Business Students.* Pearson.

Schrauf, R., & Navarro, E. (2005). Using existing tests and scales in the field. *Field Methods,* 373-393.

Shirer, M. (2022, September 12). *IDC Global*. Retrieved from IDC Global. Retrieved from Worldwide Spending on AI-Centric Systems Will Pass $300 Billion by 2026,: https://www.idc.com/getdoc.jsp?containerId=prUS49670322

Simonton, D. (2017). Big-C versus little-c creativity: Definitions, implications, and inherent educational contradictions. *Creative contradictions in education: Cross disciplinary paradoxes and perspectives*, 3-19.

Simonton, D. K. (2009). Varieties of (scientific) creativity: A hierarchical model of domain-specific disposition, development, and achievement.

Stein, M. I. (1953). Creativity and culture. *Journal of Psychology, 36,* 311-322.

Steiner, G. A. (1965). The creative organization. *Chicago: University of Chicago Press., 10*.

Sternberg, R. J., & Kaufman, R. C. (1999). *Handbook of Creativity.* Cambridge University Press.

Taylor, A. &. (2006). Superman or the Fantastic Four? Knowledge combination and experience in inno? vative teams. *Academy of Management Journal, 49*, 723 740.

Teijeiro, M. R. (2013). Graduate competencies and employability: The impact of matching firms' needs and personal attainments. *Economics of Education Review, 34,* 286-295.

Tucker, M. L., & McCarthy, A. M. (2001). Presentation self-efficacy: Increasing communication skills through service-learning. *Journal of Managerial Issues*, 227-244.

Utdanning.no. (2023, 10 23). *Utdanning.no* . Retrieved from Utdanning.no - Offentlig og kvalitetssikret. Retrieved from Studiekalkulatoren: https://utdanning.no/karakterkalkulator/studievelger

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked 18 priming: A meta-analysis. *Psychological Bulletin, 135*(3), 452-477.

Vărzaru, A. A. (2022). Assessing artificial intelligence technology acceptance in managerial accounting. *Electronics, 11(14)*, 2256.

Venkatesh, V. M. (2000). A longitudinal field investigation of gender differences in individual technology adoption decision-making processes. *Organizational behavior and human decision processes, 83(1)*, 33-60.

Wadhwa, A. &. (2006). Knowledge creation through external venturing: Evidence from the telecommunications equipment manufacturing industry. *Academy of Management journal, 49(4)*, 819-835.

Wan, Q., Hu, S.-Y., Zhang, Y., Wang, P.-H., & Bo Wen, a. Z. (2023). It Felt Like Having a Second Mind': Investigating Human-AI Co-creativity in Prewriting with Large Language Models. *ArXiv 2307.10811*.

Wang, B. R. (2022). Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. *Behaviour & information technology, 42(9)*, 1324 - 1337.

Wang, P. (2019). On defining artificial intelligence. *Journal of Artificial General Intelligence, 10*(2), 1-37.

Wolfram, S. (2023, February 14). *Stephenwolfram.com*. Retrieved from https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work./

Wu, Z., Ji, D., Yu, K., Zeng, X., Wu, D., & Shidujaman, M. (2021). AI creativity and the human-AI co-creation model. *uman-Computer Interaction. Theory, Methods and Tools: Thematic Area, HCI 2021. Springer International Publishing*, 171-190.

Zeng, L., Proctor, R. W., & Salvendy, G. (2009). Fostering creativity in service development: Facilitating service innovation by the creative cognition approach. *Service Science, 1*, 142-153.

Zoller, U. (1999). Research Based Evaluation in HOCS-Oriented Science Education and Curriculum Development. *Research in Science Education in Europe*, 183 - 191.

ZOLLER, U. (2(1)). Alternative assessment as (critical) means of facilitating HOCS-promoting teaching and learning in chemistry education. *Chemistry education Research and practice*, 9-17.

Zoller, U. (2003). HOCS Problem Solving vs. LOCS Exercise Solving: What Do College Science Students Prefer? *In Science Education Research in the Knowledge-Based Society*, pp. 201-207.

8. Appendix

8.1 Survey

The number, such as “1. Survey Start “ indicates the order that the pages appear in. The names “Control Group”, “Treatment Group 1” and “Treatment Group 2” appear to indicate that the following picture is how the group received the page.

1. Survey Start

NHH

Forskningsprosjekt

Les det følgende nøye før du samtykker:

Hvem er ansvarlig for forskningsprosjektet?
Norges Handelshøyskole (NHH) er ansvarlig for prosjektet. Prosjektet ledes av masterstudentene Anders Pay Eriksen og Aleksander Stabell Eriksen, med veileder Eirik Sjåholm Knudsen ved Institutt for Strategi og Ledelse.

Ditt personvern
Alle svarene vil være anonymisert. Det er også frivillig å være med på prosjektet. Eposten som samles inn frivillig etter eksperimentet blir slettet etter at premier er trukket.

Hva skjer med svarene når vi avslutter forskningsprosjektet?
Prosjektet skal etter planen avsluttes innen 20.12.2023. Alle anonyme data som er samlet inn vil bli arkivert for etterprøvbarhet og videre forskning. Det vil kun være prosjekteierne og forskningsmiljøet ved Norges Handelshøyskole som vil ha tilgang på dataen.

Hvor kan du finne ut mer?
Hvis du har spørsmål til studien, ta kontakt med:
Aleksander Stabell Eriksen - aleksandereriksen7@gmail.com eller 47902585
Anders Pay Eriksen - anderspayeriksen@hotmail.no eller 95994046

Ved andre henveldelser eller spørsmål knyttet til dine rettigheter, ta kontakt med:
Norges Handelshøyskole ved Eirik Sjåholm Knudsen, Eirik.Knudsen@nhh.no.

Samtykkeskjema
Jeg har mottatt og forstått informasjon om eksperimentet.

Jeg samtykker til:
- å delta på nettbasert ekspriment
- å svare på spørreskjema
- at mine svar lagres etter prosjektslutt, for etterprøvbarhet og videre forskning

Jeg samtykker

→

2. Experiment Start - (All)

NHH

Eksperiment:
Vi ber om at eksperimentet gjennomføres ved hjelp av datamaskin. Forsikre deg om at du ikke blir forstyrret eller distrahert under eksprimentet. Det er viktig at du leser intruksene nøye og følger dem under hele eksprimentet.

Neste

3. Prompt - (Treatment Group 2)

NHH

I dette eksperimentet skal du komme frem til en forretningside gjennom interaktivt bruk av generativ AI. Her følger noen tips til hvordan utnytte generativ AI best mulig.

Optimal bruk av generativ AI
Til tross for (eller faktisk på grunn av) alle dens begrensninger, er generativ AI (f.eks. ChatGPT) godt rigget for idégenerering. Man trenger ofte mange idéer for å få gode idéer, og generativ AI er god på generere store volum. Ved å stille de riktige spørsmålene, tvinger du den til å være svært kreativ.
Tips for god interaksjon med generativ AI:

1. Be ChatBoten om å generere flere idéer (Bruker her fordelen med at den er god på volum): De første idéene ChatGPT generer er ofte veldig like. Ved å spørre om flere ekspempler vil svarene bli mindre generiske. Hvis du deretter følger en idé, og spør om flere lignende, vil ChatGPT gi deg mer kreative svar.

2. Spør Chatboten om den kan være kreativ: Hvis du ønsker kreative svar, er det bare å spørre Chatboten om det. Du vil da få mer kreative idéer, som du kan forfølge gjennom samtale og etterhvert forme til å bli en forettningsidé.

3. Gi Chatboten en klar instruksjon: En god idé er å legge ved den konkrete problemstillingen du står ovenfor. Dette vil gjøre at generativ AI har bedre muligheter til å hjelpe deg med akkurat det du ønsker svar på.

4. Still oppfølgningsspørsmål: Dette vil hjelpe med å forbedre produktet, utforske muligheter og kan føre til et produkt som treffer målgruppen bedre og har større muligheter i markedet.

5. Du spør om den kan dekke et bredt spekter av forretningsaspekter: Du kan spørre språkmodellen om forskjellige aspekter som for eksempel verdiforslag, inntekts- og kostnadssiden, og markedsføringsplattformer.

Har du forstått hvordan man bruker generativ AI optimalt?

Ja

4. Experiment intro (All)

Control Group

NHH

Eksperimentinnføring:
Du skal nå gjennomføre en oppgave som skal løses innen **15 minutter**. Hvis du ikke leverer svaret innen femten minutter, blir det foreløpige svaret lagret automatisk. Du vil få all relevant informasjon om eksperimentet i neste vindu.

Vi er kun interessert i dine egne ideer i dette forsøket, og det er derfor viktig at du ikke **kommuniserer med andre eller bruker noen form for hjelpemidler under eksprimentet.**

Når du er sikker på at du har forstått oppgaven, starter du eksprimentet ved å trykke på "pilen".

→

Treatment Group 1 & 2

NHH

Eksperimentinnføring:
Du skal nå gjennomføre en oppgave som skal løses innen **15 minutter**. Hvis du ikke leverer svaret innen femten minutter, blir det foreløpige svaret lagret automatisk. Du vil få all relevant informasjon om eksperimentet i neste vindu.

Du må kun bruke generativ AI som hjelpemidel, via linken som ligger under:
Link til Chatbot **(Chatboten vil åpne i ny fane)**

Dette programmet fungerer på lik linje som ChatGPT 4 og det er mulig å holde en samtale med AI'en.

Når du er sikker på at du har forstått oppgaven, starter du eksprimentet ved å trykke på "pilen".

→

5. The Experiment (All)

Control Group

NHH

Timing

These page timer metrics will not be displayed to the recipient.

First Click	0 seconds
Last Click	0 seconds
Page Submit	0 seconds
Click Count	0 clicks

1457

Beskrivelse:
Du er en gründer som ønsker å generere en **kreativ ny forretningsidé**. Forretningsidéen (produktet) er rettet mot **studenter i Norge**. Produktet kan være et fysisk produkt, en tjeneste, en programvare, eller noe annet. Uavhengig av prismodell skal produktet være i en prisklasse som gjør det **økonomisk tilgjengelig for gjennomsnittsstudenten** i Norge. Produktet trenger ikke å eksistere ennå. Du skal komme opp med **én ide**. Idéen skal forklares med maks **850 tegn (150 ord), med en tydelig overskrift**. Skriv som om du skal forklare idéen til en potensiell investor.

Minimumstiden er 2 minutter, pilen videre blir da synlig.

Characters remaining: 850

Treatment Group 1 & 2

NHH

Timing

These page timer metrics will not be displayed to the recipient.

First Click	0 seconds
Last Click	0 seconds
Page Submit	0 seconds
Click Count	0 clicks

1455

Du må som nevnt bruke generativ AI som hjelpemidel, via linken som ligger under:
Link til Chatbot **(Chatboten vil åpne i ny fane)**

Beskrivelse:
Du er en gründer som ønsker å generere en **kreativ ny forretningsidé.** Forretningsidéen (produktet) er rettet mot **studenter i Norge**. Produktet kan være et fysisk produkt, en tjeneste, en programvare, eller noe annet. Uavhengig av prismodell skal produktet være i en prisklasse som gjør det **økonomisk tilgjengelig for gjennomsnittsstudenten** i Norge. Produktet trenger ikke å eksistere ennå. Du skal komme opp med **én ide**. Idéen skal forklares med maks **850 tegn (ca. 150) ord, med en tydelig overskrift**. Skriv som om du skal forklare idéen til en potensiell investor.

Minimumstiden er 2 minutter, pilen videre til da synlig.

Characters remaining: 850

6. Survey Start - (All)

NHH

Ideen din er nå registert! **Vi trenger litt mer informasjon om deg før eksperimentet avsluttes. Resten av undersøkelsen tar omkring 2 minutter.**

Neste

7. Questions on AI and Technology - (Treatment Group 1 & 2)

NHH

Når var første gang du brukte generativ AI? (Generativ AI kan for eksempel være ChatGPT, Midjourney osv)

- For mer enn 1 år siden
- November 2022 - Februar 2023
- Mars 2023 - Juni 2023
- Juli 2023 - Oktober 2023
- November 2023 - Februar 2024
- Har ikke brukt generativ AI

Hvor erfaren er du med ChatGPT, eller tilsvarende generative AI verktøy?

- Ingen erfaring
- Lite erfaring
- Noe erfaring
- Mye erfaring
- Veldig mye erfaring

Hvor enig er du i utsagnet: "Jeg prøver ofte ut ny teknologi"

- Sterkt uenig
- Litt uenig
- Hverken enig eller uenig
- Litt enig
- Sterkt enig

8. Creativity - (All)

First part

NHH

Nedenfor følger påstander som muligens kan gjelde for deg. Vennligst velg hvor enig eller uenig du er i påstandene: **(Jeg...)**

Er orginal og finner på nye ideer

- Sterkt uenig
- Litt uenig
- Hverken enig eller uenig
- Litt enig
- Sterkt enig

Er nysgjerrig på mange forskjellige ting

- Sterkt uenig
- Litt uenig
- Hverken enig eller uenig
- Litt enig
- Sterkt enig

Har en aktiv fantasi

- Sterkt uenig
- Litt uenig
- Hverken enig eller uenig
- Litt enig
- Sterkt enig

7(2). Second part (All)

Liker å reflektere, leke med ideer

- Sterkt uenig
- Litt uenig
- Hverken enig eller uenig
- Litt enig
- Sterkt enig

Er oppfinnsom

- Sterkt uenig
- Litt uenig
- Hverken enig eller uenig
- Litt enig
- Sterkt enig

Neste

9. Personal Information (All)

NHH

Hvor gammel er du? *

- 17 - 20
- 21 - 23
- 24 - 26
- 27 eller eldre

Hvilket kjønn identifiserer du deg som? *

- Mann
- Kvinne
- Annet

Neste

9. Field of Study - (All)

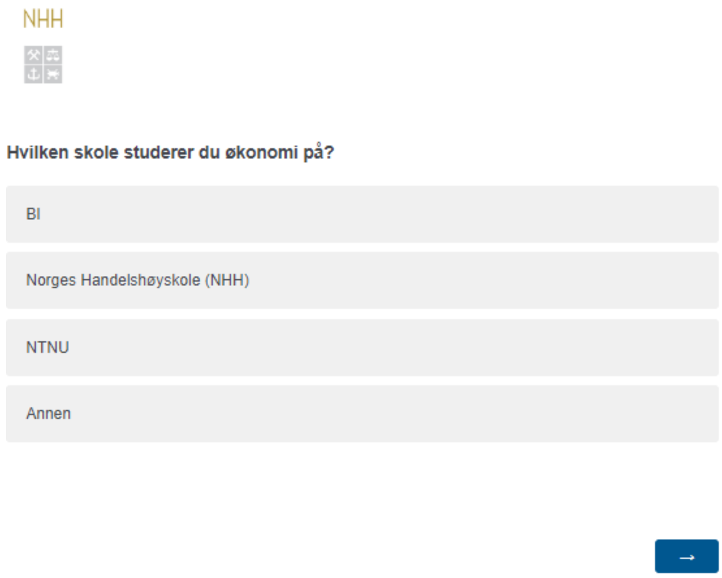
NHH

Hvilken studieretning går du?

- Medisin, ontologi, helse- og sosialfag
- Rettsvitenskap (jus)
- Samfunnsfag, psykologi, lektor- eller lærerutdanning
- Teknologi, ingeniørfag og arkitetur
- Økonomi, marked og administrasjon
- Annen studieretning

→

9(2). If Business is chosen - what school (All)



NHH

Hvilken skole studerer du økonomi på?

BI

Norges Handelshøyskole (NHH)

NTNU

Annen

→

10. GPA and Bachelor or Master degree - (All)

NHH

Estimert ditt snitt

(Estimer omtrentlig) Bachleorsnitt for bachleorstudenter og mastersnitt for masterstudenter.

A (5 - 4.5)
B (4.4 - 3.5)
C (3.4 - 2.5)
D (2.4 - 1.5)
E / F (1.4 - 0)

0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

Snitt

Er du for øyeblikket på bachleor- eller mastergrad?

Bachleor

Master

11. Length of Study (All)

Hvor lenge har du studert på nåværende studieretning?

- 1
- 2
- 3
- 4
- 5 eller høyere

Har du studert tidligere, isåfall, hvor mange år?

- Har ikke studert tidligere
- 1
- 2
- 3
- 4
- 5 eller flere

Har du tidligere arbeidserfaring (innenfor økonomi)?

- Ja
- Nei

→

12. Redirection to Forms

Tusen takk for at du tok deg tid til å svare på eksperimentet!

For å vinne premier må du legge ved e-posten din under. E-posten vil ikke kunne kobles mot eksperimentbesvarelsen din.
Premier:

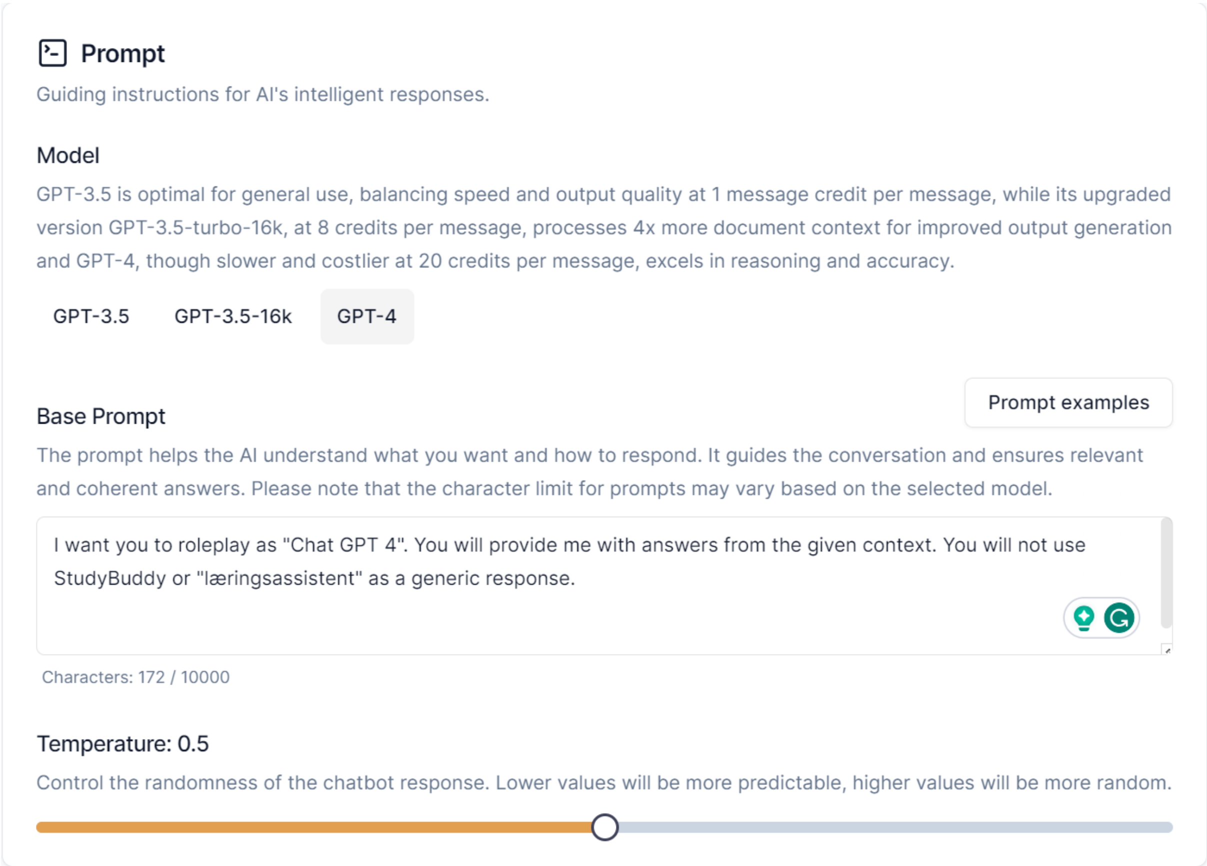
2 gavekort på 5000kr fra Norrøna

2 gavekort på 5000kr fra DBjourney (tidligere Douchebags)

Premiene blir delt ut tilfeldig blant deltakerne, ikke basert på svar på eksperimentet. Premiene blir delt ut i starten av desember.

8.2 GPTtrainer - Our Chatbot

A2.1 Prompt and temperature



A2.2 Startscreeen for our chatbot

8.3 The Judges

A2.1 The list of judges

Name	Relevant Experience
Martin Schutt	Co-Founder & Managing Partner Askeladden
Kjartan Kalstad	Partner McKinsey, lead McKinsey Digital Nordics & Leap Academy Globally
Jan Grønbech	Previous CEO Google Nordics
Alexander Haneng	Director of Innovation & Sustainability NHO Service og Handel
Ingrid Lorange	CEO Gjensidige Stiftelsen, Previous CEO Siva
Ingrid Teigland	Managing Partner Headen Ventures, Board Member Invest Europe
Johan Gjesdahl	Managing Partner Alliance Ventures
Rolv-Erik Spilling	Managing Partner Scale Leap Capital, Previous CEO Telenor Digital Services
Gunnar Sellæg	Founder Partner Core Eqty, Previous CPO Telenor
Pirasant Premraj	Supporting CEO Investment Strategy Antler
Pål Brynsrud	Chairman & Partner Credo Partners
Henrik Lisæth	Co-Founder & Partner Explore Equity, Chairman Skagen Funds
Torleif Ahlsand	Co-Founder & Parner Northzone Venture Capital
Mads Agerup	Managing Partner Industrifinans
Ole Larsen	COO & Head of Strategy Dealflow
Truls Braataas	Founder & Co-CEO DBjourney
Johan L. S. Karlsen	Professor NHH Entrepreneurial Finance
Klaus Røiri	Partner Innoco, Chairman Innolab
Stine Sofie Grindheim	CEO Dealflow, Previous Head of Digital Innovation DNB Markets
Åshild Fossum	Lead Innovation Sprint Consulting
Daniel Sørli	Founder & CMO Dr.Dropin
Per Kristian Tandberg	Managing Director Global Data Resources
Knut Glad	Managing Partner Føyen
Thomas Berglund	Founder & CEO Ocular
Per Christian Johanessen	CEO Centra Group, member of the Ethics Committee for NSA
Kristian Næss	Investmest manager Arkwright X
Torkel Muri	Sales Director Telenor Linx, Co-Founder Talkmore
Alex Adams	Co-Founder & CEO Varsity
Lars Flesland	Head StartupLab Bergen

Table 13 - The judges

A2.2 - Information for the Judges

Dommerinformasjon

Tusen takk for at du stiller som dommer i eksperimentet vårt!

Du nå skal nå vurdere 52 forretningsidéer. Forretningsidéene er på maksimalt 150 ord og hver idé er generert på mellom 2 og 15 minutter. Testdommerne våre brukte i igjennomsnitt mellom 60 og 90 sekunder på å vurdere en idé, men bruk den tiden du mener er nødvendig for å gi en representativ score.

Veiledende vurderingskriterier er vedlagt. Vi ber om at dere benytter disse og vurderer idéene på hver av de fire dimensjonene. Når det er sagt er vi ute etter deres subjektive vurdering av hver idé, basert på deres egne erfaringer og forretningsforståelse. Vi er ute etter kvaliteten på idéene vurdert utifra de fire dimensjonene. Dere trenger derfor ikke bry dere om formelle "feil", for eksempel at en idé mangler en tydelig overskrift.

Dersom dere ikke anser det som nødvendig å benytte hele vurderingsskalaen trenger dere ikke gjøre det. Dere kan også gå tilbake å vurdere idéer på nytt, dersom dere får et endret syn på idéen etter å ha lest gjennom andre idéer. All informasjon du trenger skal være tilgjengelig i dette Excel-dokumentet, men vurderingskriteriene er ***også*** lagt til som en separat PDF.

Idéene skal sendes tilbake på e-post iløpet av mandag 13. november.

Ved spørsmål, ikke nøl med å ta kontakt!

Aleksander kan nås på: 47 90 25 85

The information above was included in the Excel document containing the ideas distributed to the judges. Translated into English, the passage reads as follows:

Thank you very much for serving as a judge in our experiment!

You will now evaluate 52 business ideas. The business ideas are a maximum of 150 words each, and each idea was generated in between 2 and 15 minutes. Our test judges typically took between 60 and 90 seconds to evaluate an idea, but please take the time you feel is necessary to give a representative score.

Guiding assessment criteria are attached. We ask that you use these and evaluate the ideas on each of the four dimensions. That said, we are after your subjective assessment of each idea based on your own experiences and business expertise. We are after the quality of the ideas assessed from the four dimensions. Therefore, you do not need to worry about formal "errors", such as an idea lacking a clear headline.

If you do not consider using the entire assessment scale necessary, you do not have to. You may also go back and re-evaluate ideas if you have a changed view of the idea after reading through other ideas. All the information you need should be available in this Excel document, but the assessment criteria are also added as a separate PDF.

The ideas should be sent back by email by Monday, November 13th.

If you have any questions, do not hesitate to contact us!

Aleksander can be reached at 47 90 25 85

Anders can be reached at 95 99 40 46

A2.3 - Scoring Criteria

Vurderingskriterier

De følgende kriteriene er kun veiledende og det oppfordres til å gjøre individuelle vurderinger. Vi oppfordrer til at dommerne bruker sin subjektive forretningsforståelse til vurderingen av ideene.

Alle ideer vil bli vurdert med utgangspunkt i fire ulike vurderingsmomenter. Disse fire vurderes hver for seg.

1. ***Nyskapende***: I hvilken grad ideen er original og unik, i form av at idéen ikke har eksistert i en liknende form tidligere.
2. **Nyttig:** I hvilken grad idéen løser et problem i den virkelige verden, hvor gjennomførbar den er og hvor forståelig den er.
3. **Lønnsomhetspotensiale:** I hvilken grad idéen har et lønnsomhetspotensial, sett fra et investeringsperspektiv.
4. **Treffer målgruppen:** I hvilken grad produktet er tilgjengelig for «gjennomsnittsstudenten i Norge».

Vurdering	Nyskapende	Nyttig	Lønnsomhets-potensiale	Treffer Målgruppen
1 - 2	Det er enten ingen forretningsidé eller deltakeren har kommet frem til en idé uten å eksplisitt forklare den.	Forretningsideen er ikke nyttig.	Produktet har ikke lønnsomhetspotensiale.	Produktet treffer ikke den gjennomsnittlige studenten.
3 - 4	Forretningsidéen er lite nyskapende.	Forretningsideen er lite nyttig.	Produktet har lite lønnsomhetspotensiale.	Produktet treffer i liten grad den den gjennomsnittlige studenten.
5 - 6	Forretningsidéen er relativt nyskapende.	Forretningsideen er relativt nyttig.	Produktet har noe lønnsomhetspotensiale.	Produktet treffer i noen grad den gjennomsnittlige studenten.
7 - 8	Forretningsidéen er nyskapende.	Forretningsideen er nyttig.	Produktet har lønnsomhetspotensiale.	Produktet treffer den gjennomsnittlige studenten.
9 - 10	Forretningsidéen er svært nyskapende.	Forretningsideen er svært nyttig.	Produktet har stort lønnsomhetspotensiale.	Produktet treffer i stor grad den gjennomsnittlige studenten.

The scoring criteria above was included in the Excel document containing the ideas distributed to the judges. Translated into English, the passage reads as follows:

The following criteria are only guidelines, and individual assessments are encouraged.

We encourage the judges to use their subjective business experience in the evaluation of the ideas.

All ideas will be evaluated based on four different assessment dimensions. These four are assessed separately.

Innovative*: The degree to which the idea is original and unique, in terms that the idea has not existed in a similar form before.*

Useful*: The degree to which the idea solves a problem in the real world, its feasibility, and its comprehensibility.*

Profitability Potential*: The degree to which the idea has a profitability potential, seen from an investment perspective.*

Target Group Reach*: The degree to which the product is accessible to the "average student in Norway".*

Score	Novel	Useful	Profitability Potential	Target Group Reach
1 - 2	There is no business idea or the participant does not explain the idea.	The business idea is not useful at all.	The product does not have profitability potential at all.	The product does not reach its intended target group at all.
3 - 4	The business idea is far from novel.	The business idea is far from useful.	The product has little profitability potential.	The product struggles to reach its intended target group.
5 - 6	The business idea is somewhat novel.	The business idea is somewhat useful.	The product has some profitability potential.	The product somewhat reaches its intended target group.
7 - 8	The business idea is novel.	The business idea is useful.	The product has decent profitability potential.	The product reaches its intended target group.
9 - 10	The business idea is really novel.	The business idea is really useful.	The product has excellent profitability potential.	The product is excellent at reaching its intended target group.

8.4 Variables

The list of all variables included in initial testing and final analyses:

Variables Overview	
Independent Variables:	
NoAI:	A binary variable that is 1 when the participant belongs to the Control Group and 0 otherwise.
NoPromptAI_D:	A binary variable that is 1 when the participant belongs to Treatment Group 1 and 0 otherwise.
PromptAI_D:	A binary variable that is 1 when the participant belongs to Treatment Group 2 and 0 otherwise.
AI_D:	A binary variable that is 1 when the participant belongs to either Treatment Group and 0 otherwise.
Dependent Variables:	
Novelty:	The novelty score based on evaluation of three judges, on a scale of 1-10.
Useful:	The useful score based on evaluation of three judges, on a scale of 1-10.
ProfitPotential:	The profit potential score based on evaluation of three judges, on a scale of 1-10.
TargetGroup:	The reaching the intended target group score based on evaluation of three judges, on a scale of 1-10.
Creativity:	A combination of the Novelty and Useful scores. Novelty and useful both make up 0.5 of the Creativity score.
Profit:	A combination of the ProfitPotential and TargetGroup scores. ProfitPotential makes up 0.75 and TargetGroup 0.25 of the Profit score.
TotalScore	A combination of the Creativity and Profit scores. Creativity and Profit both make up 0.5 of the TotalScore.

Control and Predictor Variables:	
Universal:	
FirstClick:	Participants' time when first clicking in the experiment, measured in seconds.
TimeOfLastEdit:	Participants' time when clicking in the experiment for the last time, measured in seconds.
TimeOfIdeaSubmit:	Participants' time when submitting their idea, measured in seconds.
NumberOfClicks:	The total number of clicks the participants had in the experiment.
SelfReportedCreativity:	A combination of 5 items taken from the "Openness to new experiences" personality trait from the Big Five, meant for predicting creative behavior.
AgeInterval:	A value of 1-4 dependent on participants' age interval where 1 represents the youngest and 4 the oldest participants.
Gender:	A binary variable that is 0 when the participant is female and 1 if the participant is male.
Gender Other:	A binary variable that is 1 if the participant does not identify as male or female and 0 otherwise.
LineOfStudy:	Categorizes peopel into six lines of studies.
LenghtOfCurrentStudy:	The number of years participants have spent on the degree they currently pursue.
YearsOnEarlierStudy:	The number of years participants have spent on previous degree(s).
TotalYearsStudying:	The total number of years participants have spent studying across all degrees.
WorkExperience:	A binary variable that is 1 if the participants have proffesional experience in business and 0 otherwise.

AI only:	
FirstTimeUsingAI:	A value of 1-6 dependent on what time interval the participants used generative AI for the first time, where 6 represents more than one year ago and 1 represents "never used genAI".
ExperienceAI:	A value of 1-5 dependent on self reported experience with generative AI, where 1 represents no experience and 6 represents highly experienced.
NewTechnology:	A value of 1-5 dependent on self reported adoption rate of new technology products, where 1 represents low and 6 high adpotion rate.
TotalInteractions:	The number of messages sent back and forth between the chatbot and the partcipant.
UserResponses:	The number of messages the participant sent to the chatbot.
One_Response_D:	A binary variable that is 1 if the participant sent only one message to the chatbot and 0 otherwise.
DirectCopy_D:	A binary variable that is 1 if the participant directly copied its final response from the chatbot and 0 otherwise.
DirectCopy_OneResponse_D:	A binary variable that is 1 if the participant only sent one message to the chatbot and directly copied its final response from the chatbot, and 0 otherwise.
Moderation Variables:	
TimeStudying_Moderation	The moderation effect total time spent studying have on AI's effect on creativity.
Grade_Moderation	The moderation effect GPA have on AI's effect on creativity.

Table 14 - Explanation of variables

8.5 T-tests

Overview of all conducted t-tests for all seven dependent variables, comparing all three participant groups:

t-tests

	Hypothesis	t	df	p
Novel	noAI < AI	-1,53	394	0,064
	noAI < AIuP	-0,44	259	0,33
	noAI < AImP	-2,25	255	**0,013**
	AIuP < AImP	-1,91	272	**0,029**
Useful	noAI > AI	-0,49	394	0,686
	noAI > AIuP	-0,66	259	0,744
	noAI > AImP	-0,37	256	0,645
	AIuP < AImP	-0,25	273	0,4
ProfitPotential	noAI > AI	-0,6	394	0,725
	noAI > AIuP	-0,62	259	0,734
	noAI > AImP	-0,39	255	0,653
	AIuP < AImP	-0,29	272	0,388
TargetGroup	noAI > AI	-0,61	394	0,73
	noAI > AIuP	-0,54	259	0,704
	noAI > AImP	-0,51	255	0,696
	AIuP < AImP	-0,02	272	0,492

t-tests

	Hypothesis	t	df	p
Creativity	noAI < AI	-0,55	394	0,29
	noAI > AIuP	-0,14	259	0,556
	noAI < AImP	-1,1	255	0,136
	AIuP < AImP	-1,31	272	0,096
Profitabilty	noAI > AI	-0,66	394	0,745
	noAI > AIuP	-0,66	259	0,745
	noAI > AImP	-0,47	255	0,679
	AIuP < AImP	-0,24	272	0,406
Total score	noAI < AI	0,05	394	0,519
	noAI > AIuP	-0,43	259	0,667
	noAI < AImP	-0,36	255	0,358
	AIuP < AImP	-0,84	272	0,2
Time spent	noAI > AIuP	4,82	259	**<.001**
	noAI > AImP	0,53	255	0,299
	AIuP < AImP	-4,11	272	**<.001**
UserResponses	AIuP < AImP	-5,73	272	**<.001**
One_Response	AIuP > AImP	3,77	272	**<.001**

Table 15 - T-tests

8.6 Assumptions for OLS

Testing assumptions for OLS on the whole population

Plotting the residuals versus the fitted values in the scatter lot allows us to inpect whether our dataset has a constant variance, detecting heteroscedasticity. Though there are some concentrations of residuals around 3.5 to 4.0, the residuals are mainly evenly distributed around the red line. Subsequently, we do not consider heteroscedasticity a noteworthy concern. Moreover, as the residuals somewhat resemble a straight line, we consider the linearity assumption to be met.

Further, inspecting the Q-Q plot, the populations seem to resemble the normal distribution as the quantiles are mainly plotted on a straight line. Despite some skew in both ends, we consider the plot satisfactory for our purposes.

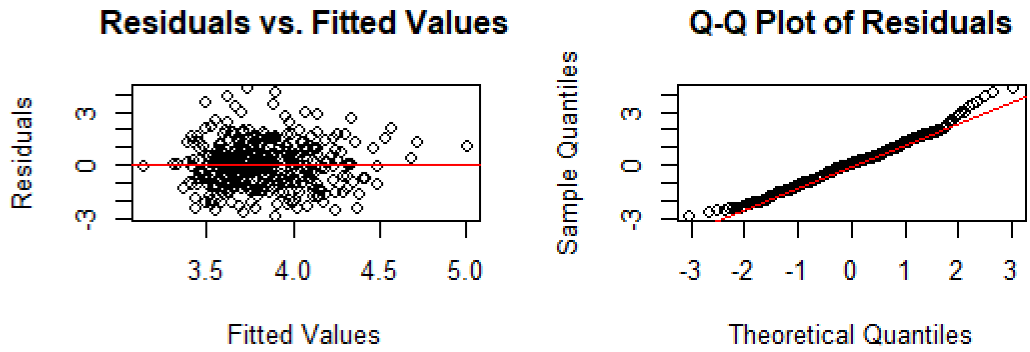


Figure 9 - OLS assumptions, whole population

To test for autocorrelation by conducting a Durbin-Watson test. As the DW-statistic is close to the desired value of 2, and the p-value is far from < 0.05 (Saunders et al., 2019), we can conclude that there is close to zero autocorrelation in the residuals.

Durbin-Watson test	
DW-Statistic	1.995
p-value	0.415

Table 16 - Durbin-Watson test, whole population

Multicollinearity was tested by calculating the collinearity statistics for tolerance and VIF. As none of the included variables achieve a tolerance above 0.1 and a VIF below 10, we conclude that there is no multicollinearity in our data.

Whole Population

Variable	Tolerance	VIF
TimeOfLastEdit	0,93	1,08
SelfReportedCreativity	0,91	1,09
AgeInterval	0,55	1,8
Gender 1	0,86	1,16
Gender Annet	0,97	1,03
GradeAverage	0,94	1,07
TotalYearsStudying	0,52	1,91
WorkExperience 1	0,88	1,14
PromptAI_D 0	0,71	1,4
NoPromptAI_D 1	0,67	1,49

Table 17 - Multicollinearity, whole population

Testing assumptions for OLS on the AI population

All arguments above are also valid when testing the assumptions for OLS on the AI population. Note that the test below are done with all control variables, and not only the AI specific variables.

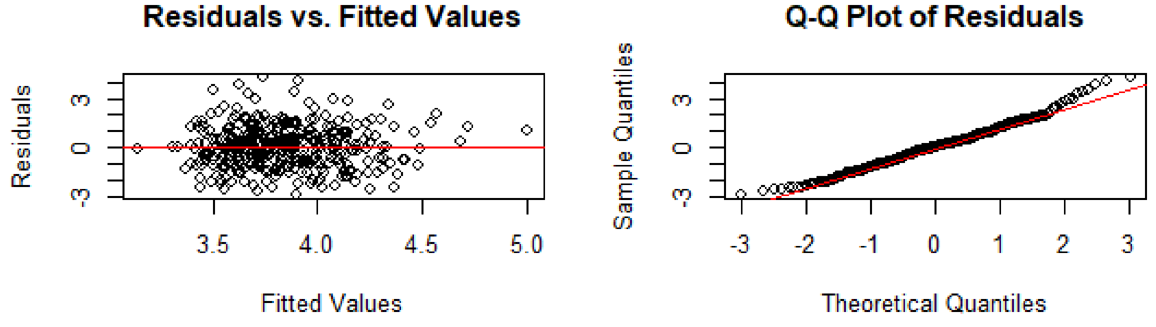


Figure 10 - OLS assumptions, AI population

Durbin-Watson test	
DW-Statistic	1.992
p-value	0.452

Table 18 - Durbin-Watson test, AI population

AI Population

Variable	Tolerance	VIF
TimeOfLastEdit	0,91	1,1
FirstTimeUsingAI	0,69	1,45
ExperienceAI	0,61	1,64
NewTechnology	0,68	1,48
SelfReportedCreativity	0,84	1,19
AgeInterval	0,56	1,8
Gender 1	0,78	1,28
Gender Annet	0,93	1,07
GradeAverage	0,89	1,13
TotalYearsStudying	0,52	1,93
WorkExperience 1	0,87	1,15
PromptAI_D 0	0,9	1,11

Table 19 - Multicollinearity, AI population

8.7 GPTrater Input

Prompt GPTrater-1

GPTrater-1 was prompted with the instructions below before evaluating ideas. Note that GPTrater-1 was called GPT Vurderinger and that "Treningsmodell_GPTrating" was the 25 ideas given to the chatbot as training.

As GPT Vurderinger, my primary role is to assess startup ideas based on Novelty and Usefulness, each rated on a scale from 1 to 10, mirroring the scores provided in "Treningsmodell_GPTrating". When similar ideas are presented, they will receive lower ratings for Novelty to reflect the lack of originality. The normal novelty score should be 3.9 and the normal usefulness score should be 4.2. I can deviate from this scores, but only if an idea is varies greatly in novelty and usefulness compared to the other ideas. Only a few ideas should score 7 or higher.

All ratings will be given as whole numbers, without decimal points. The ratings will be presented in a format suitable for exporting to two rows in an Excel sheet, with Novelty and Usefulness as separate columns. This structured approach ensures clear and efficient recording of scores, closely aligning with the ratings in "Treningsmodell_GPTrating". My evaluations will strictly include these numeric ratings, without any accompanying explanations or justifications, focusing on objective and concise assessment based on originality and practical utility.

Prompt GPTrater-2

GPTrater-2 was prompted with the following instructions before evaluating ideas:

As GPT Vurderinger, my primary role is to assess startup ideas based on Novelty and Usefulness, each rated on a scale from 1 to 10, and you must use the whole scale. Bad ideas are given bad scores, while good ideas are given good scores. The ratings will be presented in a format suitable for exporting to two rows in an Excel sheet, with Novelty and Usefulness as separate columns. This structured approach ensures clear and efficient recording of scores. My evaluations will strictly include these numeric ratings, without any accompanying *explanations or justifications, focusing on objective and concise assessment based on originality and practical utility. Use the knowledge only as a guidelines.*

8.8 Naming our Thesis

To test the impact of human-AI co-creation for ourselves, we decided to name our thesis in collaboration with ChatGPT-4. After initial prompting, we provided the chatbot with five examples of master thesis titles and our introduction. AI's first suggestion was "Synergizing Minds and Machines". After back-and-forth iteration, and many suggestions later, ChatGPT-4 came up with the title that became the name of our thesis.

You

We are two master students majoring in Business Analysis and Performance Management. We are now concluding our master thesis, writing a thesis with the following research question: "Does human-AI co-creation positively affect creativity, and does the level of competence moderate the relationship?". You are a seasoned professor who have written several acclaimed papers with the following titles; "The

ChatGPT

Based on the thorough and insightful introduction you've provided, the title of your master thesis should encapsulate the essence of your research, highlighting the intersection of human creativity, AI co-creation, and the role of competence. Here's a suggested title that captures these elements:

"Synergizing Minds and Machines: Exploring the Impact of Human-AI Co-Creation on Creativity and the Moderating Role of Competence"

This title reflects the core elements of your research – the synergy between human and AI in the creative process, the exploration of how this interaction affects creativity, and the investigation into whether individual competence levels play a significant role in this dynamic.