The More Is Not the Merrier: Effects of Prompt Engineering on the Quality of Ideas Generated By GPT-3

Lucas Memmert
Universität Hamburg
lucas.memmert@uni-hamburg.de

Izabel Cvetkovic
Universität Hamburg
izabel.cvetkovic@uni-hamburg.de

Eva Bittner
Universität Hamburg
eva.bittner@uni-hamburg.de

Abstract

Generative language models (GLM) like GPT-3 can support humans in creative tasks. Such systems are capable of generating free-text output based on a provided input prompt. Given the outputs' sensitivity to the prompt, many techniques for prompt engineering were proposed both anecdotally in social media and increasingly in literature. It is, however, unclear if and how such a system and such techniques can be employed in creative contexts such as for generating ideas. In our study, we investigate the effects of using six prompt engineering techniques. For each combination of techniques, we have GPT-3 generate ideas for an exemplary scenario. The ideas are rated according to novelty and value. We report on the effects of the (combinations of) prompt engineering techniques. With our study, we contribute to the emerging field of prompt engineering and shed light on supporting idea generation with GLMs, showing a pathway to embedded GLM capabilities.

Keywords: brainstorming, generative language model, GPT-3, human-AI collaboration, prompt engineering

1. Introduction

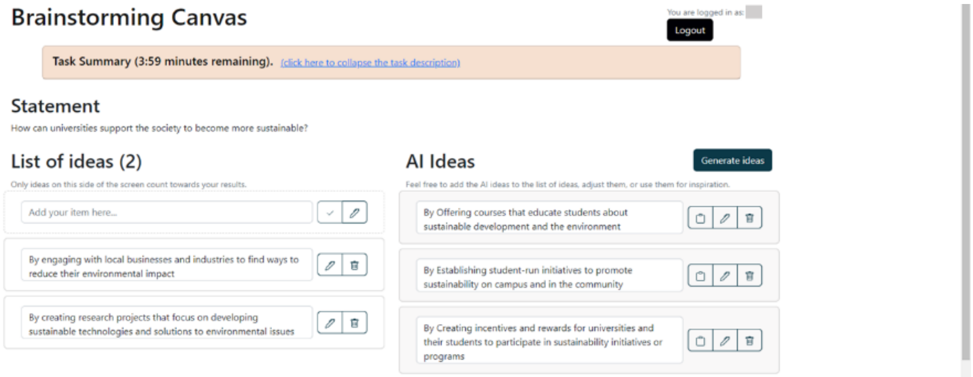
Generative language models (GLM) like GPT-3 (Brown et al., 2020) have gained increasing attention for their ability to support humans in various creative tasks. Such systems are capable of generating free-text output based on a provided free-text input, typically referred to as a *prompt*. With recent advances, users can even directly interact with such GLMs via chat (e.g., ChatGPT). While such an interaction directly with the model is possible, GLM capabilities can be embedded into products, essentially abstracting away the complexities of interacting with the GLM for the users. An example of such a system is depicted in Figure 1. In this idea generation application prototype, the user can add a question, add first ideas, and request AI ideas if desired. The user can review the suggested ideas and select or ignore them as they wish. The suggestions are generated dynamically in the background. The user does not need any knowledge on how to interact with GLM but can still benefit from its capabilities. This removes the strain on the user to formulate a *good* prompt, shifting it to the tool designer. During tool development, the *prompt template* to generate ideas needs to be defined. The template needs to be flexible to accept different questions for which the user might want to generate ideas.

The sensitivity of the GLM's output to the input (Zhao et al., 2021) raises questions about how to effectively design a *good* prompt template for ideation purposes. While prompt engineering techniques have been explored for typical natural language processing tasks (Brown et al., 2020), there is a lack of systematic investigation on the effects of these techniques in creative tasks such as ideation. Therefore, in this study, we aim to investigate the effects of six different prompt engineering techniques on the quality of ideas generated by GPT-3 on an exemplary idea generation question. To this aim, we pose the research question: *How does prompt engineering, using a GLM like GPT-3, affect the quality of ideas generated for an ideation task?*

To investigate this question, we developed a baseline prompt template and all variations of the prompt template considering the six prompt engineering techniques. We filled the prompt template with an exemplary brainstorming scenario and had GPT-3 generate ideas for each prompt. The quality of the generated ideas is assessed by three rates according to novelty and value, which are typical quality criteria for idea evaluation (Siangliulue et al., 2015). We analyzed the effects of the prompt engineering techniques and all their combinations on the idea quality. Our findings suggest that more is not necessarily merrier when it comes to prompt design for creative tasks. Combining certain prompt design techniques can negatively affect the quality of ideas generated by GPT-3.

Our study contributes to the emerging field of prompt design and aims to shed light on the potential of GLMs to support ideation. Moreover, it shows a potential path for abstracting away the complexities of

URI: https://hdl.handle.net/10125/107289
978-0-9981331-7-1



Brainstorming Canvas

Task Summary (3:59 minutes remaining). (click here to collapse the task description)

Statement

How can universities support the society to become more sustainable?

List of ideas (2)

Only ideas on this side of the screen count towards your results.

Add your item here...

By engaging with local businesses and industries to find ways to reduce their environmental impact

By creating research projects that focus on developing sustainable technologies and solutions to environmental issues

AI Ideas

Feel free to add the AI ideas to the list of ideas, adjust them, or use them for inspiration.

By Offering courses that educate students about sustainable development and the environment

By Establishing student-run initiatives to promote sustainability on campus and in the community

By Creating incentives and rewards for universities and their students to participate in sustainability initiatives or programs

Figure 1. Screenshot of a GLM-based app prototype adapted from Memmert and Tavanapour (2023)

directly interacting with GLMs while still benefiting from its capabilities, offering a wrapper UI, helping the user to ask questions to the GLM (Dang et al., 2022). Our paper offers the foundation for studies examining humans and AI brainstorming together, as it provides insight into a key design aspect: the underlying prompt template. This work informed a follow-up study in which the actual effect of providing AI suggestions to humans during a brainstorming session was assessed according to novelty and value of ideas (under review).

2. Background

2.1 Idea generation support

The goal of idea generation is to produce as many out-of-the-box ideas as possible to solve burning problems (Schallmo & Lang, 2020). To facilitate the creativity process, computer-based tools called Creativity Support Systems (CSS) have been developed. CSS has a long research history and focuses on different aspects of the creativity process (Przybilla et al., 2019). CSS can provide participants with cognitive or social stimulation (Pilcicki et al., 2022). Different tools have been developed and evaluated, and they show varying degrees of effectiveness (Gabriel et al., 2016).

One of the main areas of focus in CSS is idea generation tools, which aim to support the creative process by providing various techniques and methods for generating ideas. These tools can have a positive impact on the creative process, including improving the quality and quantity of generated ideas (Maaravi et al., 2021). However, these tools also have some drawbacks. One of the main challenges is that they often rely on a fixed set of methods and techniques, which may limit their effectiveness. Moreover, some of the tools may require a certain level of expertise or training to use them effectively (Frich et al., 2019). Recently, GLMs have emerged as a new technology that may offer a more effective way to support the creativity process.

2.2 Generative language models (GLM)

GLMs are a type of machine-learning-based AI systems. GLMs are trained on large corpora and are able to predict the next word given a certain input. GPT-3 is a powerful example of such GLMs (Brown et al., 2020). Due to their generative nature, GLMs have been proposed to be used for creative applications (Gero et al., 2022). However, there are several challenges with such systems. To interact with GLMs, an input text, i.e., a prompt, needs to be formulated. This can be difficult for novices, who might be surprised by the outputs of such systems at first use (Jiang et al., 2022). Dang et al. (2022) suggest supporting users in asking questions to the model. While recent advances allow humans to freely interact with GLMs via chat (e.g., ChatGPT), and one might educate users in effectively working with GLMs, a direct interaction might not always be feasible. In such cases, the GLM capability can be embedded into a system, hidden from the users. For this, a prompt template would need to be pre-defined at development time and populated with the user input at runtime. In this study, we investigate how such a prompt template should be formulated for an idea-generation application by exploring different techniques for phrasing prompts.

With regard to using GLMs for idea generation, Di Fede et al. (2022) suggested using GPT-3 for brainstorming but have not reported data yet. Stevenson et al. (2022) used GPT-3 (earlier version) for the "alternative use test" creativity test, generating ideas for alternative uses for an object. In this study, we focus on using GLMs for a different divergent thinking task, i.e., to generate ideas for solving societal problems.

3. Method

3.1 Prompt template development

Requirements for prompt templates. We seek to develop a prompt template that can be embedded into an idea-generation app (see Figure 1). The prompt template should be *flexible* to be used for different questions, i.e., it should accept a parameter for the specific question at hand (see Figure 2). The prompt, once entered into the GLM, should produce a *list of N ideas* (instead of a free text), which can be shown as items to the user in the app. Lastly, the prompt should produce *good ideas*. The goodness of ideas can be operationalized or measured according to different criteria. For this study, we use *novelty* and *value* of ideas with a definition adapted from Siangliulue et al. (2015), as these reflect common evaluation criteria used in brainstorming research (Althuizen & Reichel, 2016; Haase & Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023). *Novelty* will be rated by considering how *novel, original, or surprising* the idea is. *Value* will be judged by considering how *useful* and *practical* the idea sounds. We have included six prompt engineering techniques to investigate their effect on idea quality.

Baseline. First, we developed a baseline prompt. We tested if we could provide the question to the GLM directly without further input. Note that squared parentheses indicate a placeholder filled at runtime.

```
[question]
```

Prompting the GLM directly with the brainstorming question, however, resulted in inconsistent responses for certain questions (e.g., “How can generative language models be used?”). The results sometimes were lists of ideas (as intended), but sometimes free text, unfit for our app. Thus, we added a description of the task we expect the GLM to perform (Reynolds & McDonell, 2021), i.e., ‘Provide 3 ideas’. To receive a bullet-point list easily processible by our app, the technique of “itemization” might be used (Mishra, Khashabi, Baral, Choi, & Hajishirzi, 2022) by indicating a starting enumeration. In this way, the GLM “recognizes” the desired output to be a list. This template meets the first two formal requirements, as it is *flexible* to be used for different ideation questions and provides a *list of ideas*.

```
Provide 3 ideas for the question below.
[question]
1.
```

In the following, we develop prompt templates based on prompt engineering techniques, applying them to the idea generation context to assess differences in the goodness of the results. All selected prompt engineering techniques might be used in creative contexts and require adding text to the baseline prompt template described above. The techniques fundamentally differ in their dependency on the specific brainstorming question. Three of the techniques are independent of the ideation question, while the others are question-specific. As a result, to implement the former techniques, changes only occur in the backend, whereas for the latter, the users would need to enter additional information (i.e., the GUI would need to be adapted). Table 1 offers an overview of the techniques with their respective characteristics; Figure 2 shows how the type of technique affects information flow.

Table 1. Prompt engineering techniques and characteristics

Ideation question dependency	Additional User Input	Prompt Engineering Techniques
Independent	No	• Evaluation criteria specification • Instructions/Schema • Demonstrations – question independent
Dependent	Yes	• Providing context • Expert perspective • Demonstrations - question dependent

Evaluation criteria specification (ECS). Specifying expectations is an important part of prompting GLMs. For brainstorming, we seek *good* ideas according to the two dimensions novelty and value with adapted definitions from Siangliulue et al. (2015), as these include common evaluation criteria for brainstorming ideas. Explicitly including expected idea characteristics in the prompt was done before (e.g., Stevenson et al., 2022) and can improve result quality (Summers-Stay et al., 2023). Thus, we included these criteria in the prompt (difference to baseline in blue):

```
Provide 3 novel and valuable ideas for the question below. Novelty will
be rated by considering how novel, original or surprising the idea is.
Value will be judged by considering how useful and practical the idea
sounds.
[question]
1.
```

Instruction/schema (INS). Another suggested approach is to use schemas to identify the different

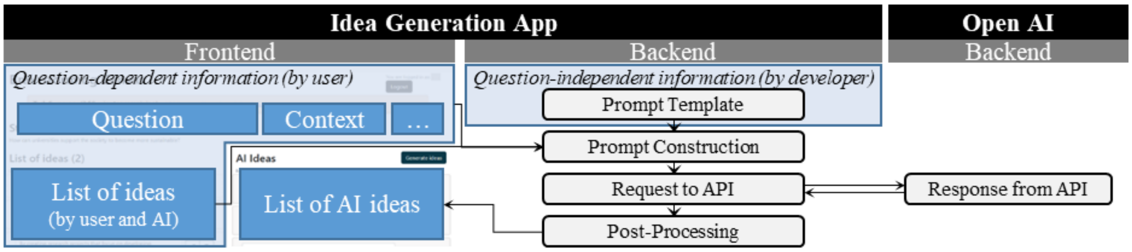


Figure 2. Illustrative depiction of the dynamic population of the prompt template with question-dependent information by the user to dynamically generate suggestions (via an API call)

prompt parts, e.g., by using headers (Liu et al., 2023; Mishra, Khashabi, Baral, & Hajishirzi, 2022) to label the task, question, and expected answer. This could be particularly relevant once more prompt techniques and, thereby, more information is added to the prompt.

```
Task: Provide 3 ideas for the question below.
Question: [question]
Answer:
1.
```

Demonstrations – question independent (DQI). Providing examples within the prompt can help to calibrate the GLM to better “understand” the expected results. This technique is commonly referred to as demonstration or few-shot learning (as opposed to zero-shot learning), builds on the idea of analogical reasoning (Q. Zhu & Luo, 2022) and was shown to be effective in some cases (Brown et al., 2020; Liu et al., 2023). For the area of creativity and design, Q. Zhu and Luo (2022) used design competition questions and winning ideas from previous years as examples to then generate new ideas. Inspired by this, we used pairs of brainstorming questions and good (scoring high for novelty and value) ideas published in prior studies (Nelissen, 2022; Siangliulue et al., 2015). We randomized the order of the selected ideas (demonstrations) to prevent sequential order effects (Zhao et al., 2021).

```
Provide 3 ideas for the question below.
What are ideas for new sports equipment products for the student market?
1. Headphones that choose which song to play aligned with your bpm (so intense music for running/ cardio and chill music for yoga/cooling down)
2. AI analytics for evaluating form from video recording
3. A device analyzing blood to indicate which supplements and vitamins are needed

Provide 3 ideas for the question below.
< question 2 – shortened here>
1. < question 2, idea 1 – shortened here >
2. < question 2, idea 2 – shortened here >
3. < question 2, idea 3 – shortened here >

Provide 3 ideas for the question below.
< question 3 – shortened here >
1. < question 3, idea 1 – shortened here >
2. < question 3, idea 2 – shortened here >
3. < question 3, idea 3 – shortened here >

Provide 3 ideas for the question below.
[question]
1.
```

Besides these prompt engineering techniques independent from the question, there are techniques depending on the specific question, such as adding context, expert perspective, and question-dependent demonstrations. While the former techniques aimed at improving the output quality independently from the input, these techniques adjust the input (i.e., would require additional input from the user and an adjustment of the GUI, e.g., by adding additional input fields).

Providing context (CON). GLMs generalize their outputs across the different contexts present in the training data. Thus, providing the *intended* context might improve the results to be tailored to the desired outputs (Liu et al., 2023). Besides using examples (few-shot learning), this can be achieved by adding additional information in a zero-shot learning (i.e., no examples) setting (Brown et al., 2020). In our tool, this could be reflected through an additional ‘context’ text field for the user to add context-related information.

```
Provide 3 ideas for the question below.
[context]
[question]
1.
```

Expert perspective (EXP). Similarly to adding additional context, asking the GLM to take a certain perspective or angle might improve the output quality. Reynolds and McDonell (2021) suggest prompting the GLM to take the perspective of a public figure or a specific role (e.g., teacher) to produce particular results; Haase and Hanel (2023) suggest including a profession. Similarly, we ask the GLM to take the role of an expert for a specific field, which is to be specified by the user.

```
Assuming you are an expert for [expert field], please provide 3 ideas for the question below.
[question]
1.
```

Demonstrations – question dependent (DQD). Previously, we explained that examples or “demonstrations” could be used to improve the GLMs' performance (few-shot learning). While the examples in the earlier technique were independent of the specific brainstorming question, once the user entered their first ideas, these ideas can be added dynamically to the prompt. For our example, we assumed the user had already added at least three ideas, and three of these ideas are included in the prompt. We randomized the idea order to prevent ordering effects (Liu et al., 2023; Zhao et al., 2021). In actual use, the best ideas might be selected algorithmically (Summers-Stay et al., 2023).

```
Provide 6 ideas for the question below.
[question]
1. [idea 1]
2. [idea 2]
3. [idea 3]
4.
```

The six selected techniques are independent of one another, allowing for any combination (i.e., all combinations are feasible). The goal of our analysis is to increase the understanding of which (combinations of) techniques lead to novel and valuable ideas. We developed a Python script to create all 64 prompt templates consistently, representing all combinations of the six prompt techniques described above (2^6 combinations from including none to including all techniques; no duplications). We recursively applied the techniques, i.e., when ‘Context’ and ‘Demonstrations – Question Independent’ were included, then context was also added to the demonstrations. Our templates are

Table 2. Exemplary scenario

Component	Text
Context	Food waste is a major issue that affects both the environment and the economy. Globally, it is estimated that about one-third of all food produced is lost or wasted each year. Food waste is a huge source of greenhouse gas emissions and wasted natural resources, and therefore – reducing food waste could help to reduce global greenhouse gas emissions, establish food security, and encourage healthy food systems.
Question	How can we reduce food waste?
Expert Perspective	sustainability and environment
3 Examples (Demonstrations – question dependent)	• Reward grocery stores that donate their near-expired food to a food-bank where people can cook this donated food • Decompose leftovers and expired food, and supply them as fertilizers to local farmers and gardeners • Create an app that allows people to offer their leftovers or extra food to those who need it

prefix prompts as the entire prompt text proceeds the expected output (Liu et al., 2023).

3.2 Idea generation

To generate ideas for our experiment, we developed an exemplary brainstorming scenario. We used the societal problem of food waste (see Table 2). Open-ended problems for which ideas can be generated with relatively common knowledge are frequently used in brainstorming studies (Y. Zhu et al., 2020, 2021); the specific societal problem of avoiding food waste was adapted from Y. Zhu et al. (2021). While the question was copied, we added the question-specific details for context and expert perspective. For the example ideas, we used ideas that scored high in novelty and value from the same prior study. We filled all 64 templates with the contents for this specific scenario. Two additional scenarios we had included originally to increase the robustness of our results we had to abandon (as will be described in the idea evaluation section).

Populating the baseline prompt template resulted in the prompt below (scenario-specific content in orange):

```
Provide 3 ideas for the question below.
How can we reduce food waste?
1.
```

We used a custom Python script to populate all templates and make the requests to the OpenAI-API. We used the most powerful model, ‘davinci-003’, at standard settings but adjusted the temperature to 0.9, as per the documentation for creative applications. We did not make any changes to the model to improve the results or adjust them to the exemplary scenario. We post-processed the results, removing any enumeration and cutting off the ideas after the first dot, as for some ideas, added explanations made the ideas very long. This resulted in 192 ideas (3 ideas per 64 prompts). No ideas were flagged by OpenAI’s content moderation.

3.3 Idea evaluation

In an approach similar to Siangliulue et al. (2015), we generated random sets of about 25 ideas in random order for each scenario and had crowd workers assess them according to novelty and value. We provided the definition for both criteria in accordance with Siangliulue et al. (2015), with *novelty* as ‘consider how novel, original or surprising the idea is’ and *value* as ‘consider how useful and practical the idea is’. We recruited participants on the Prolific platform, restricting to English-speaking individuals with a high approval rating. We aimed at 3 ratings per idea (Siangliulue et al., 2015). We included 3 randomly selected ideas all raters had to rate, enabling us to calculate inter-rater agreement. Though most participants passed the attention check items, we decided not to use the crowd-worker evaluation data due to poor data quality, i.e., poor inter-rater agreement.

Consequently, we had three blind-to-condition raters (one design thinking expert and the first two authors) independently rate all 192 ideas for one scenario on both dimensions. Having a set of judges rate all ideas is common in brainstorming research (e.g., Althuizen & Reichel, 2016; Y. Zhu et al., 2021), so is rating individual instead of sets of ideas (e.g., Haase & Hanel, 2023; Stevenson et al., 2022). We decided to include only one scenario, allowing for testing all techniques and their combinations while reducing potential fatigue from making many evaluations. Inter-rater agreement was 0.76 for novelty and 0.57 for value, or ‘excellent’ and ‘fair’ respectively (Cicchetti, 1994). Generated ideas included: “Develop software that would help shoppers calculate exact amounts of food they need depending on meal times and leftovers” (high novelty score), “Create an online platform for restaurants to share their surplus food with people who need it” (high value score), and “Implement “best by” labeling systems” (low novelty and low value score).

3.4 Data analysis

We used custom Python scripts for pre-processing, i.e., sorting the ratings for techniques and combinations. The clean data was analyzed with JASP statistics software (JASP, 2023). After performing assumption checks and confirming the normality of data distribution (Levene’s p=.747), one-way ANOVA was performed to assess if there is an effect of prompt engineering on idea quality (value and novelty). ANOVA was performed on all possible combinations of techniques, leading to 64 pairwise comparisons. To determine the direction of the effect, significant results were further inspected with Tukey's post-hoc test, which also controls for the smaller number of observations in subgroups when performing pairwise comparisons. Since the variance of the individual ratings was higher than when averaged across the raters, we performed the same analysis on a dataset with averaged scores. Lastly, we checked whether prompt length affects the quality of the output with correlation analyses. We chose Spearman's rho coefficient as the prompt length variable did not satisfy the normality assumptions.

4. Results

This study investigates the impact of using single and combined prompting techniques on idea novelty and value for a specific brainstorming question. Regarding single techniques, the results showed that the use of Evaluation Criteria Specification had a positive and significant impact on the novelty of ideas (F=18.250, p<.001***, Figure 3-a), while Context had a negative impact compared to no technique (F=5.464, p=.02*, Figure 3-b). The use of the other single techniques did not yield any statistically significant results.

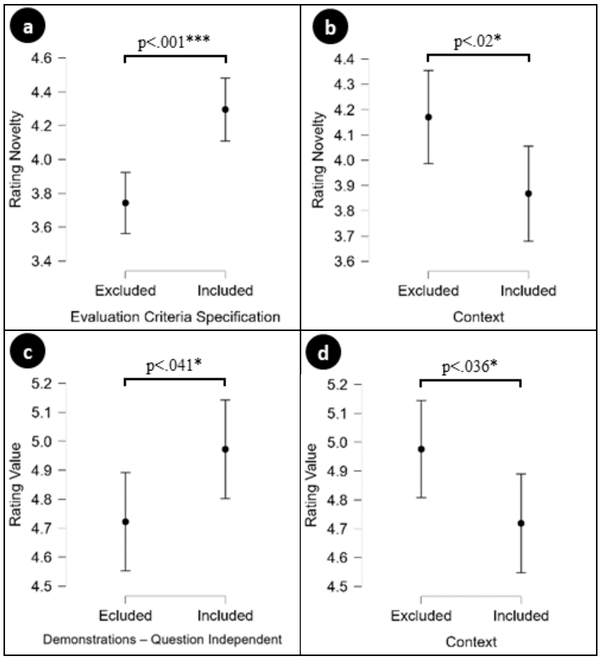


Figure 3. Techniques performing better than baseline (left) and worse than baseline (right)

Analyzing combinations of two techniques, the combination of INS and DQD showed a statistically significant effect (F=4.736, p=.03**); however, the post-hoc analysis did not reveal any significant differences between groups.

Combinations of three techniques: The combination of ESC, INS, and DQD showed a statistically significant effect (F=5.216, p=.023*), with post-hoc analysis indicating that ESC was better than INS (t=3.815, p=.004***) and DQD (t=4.460, p<.001***). ESC alone performed better than the combination of ESC and DQD (t=3.332, p=.021*) and INS and DQD (t=4.030, p=.002**). The combination of CON, DQI, and EXP showed a statistically significant effect (F=8.894, p=.003), with post-hoc analysis indicating that combining DQI with EXP leads to more novel ideas than CON alone (t=-3.439, p=.015*).

The combination of ESC, CON, DQD, and EXP showed a statistically significant effect (F=4.736, p=.03**), with post-hoc analysis indicating that CON alone performed worse than the combination of ESC and CON (t=-3.572, p=.033*). The combination of ESC and CON was also found to be better than CON and DQD (t=3.648, p=.026*).

The combination of five techniques: ESC, INS, CON, DQD, and EXP showed a significant effect (F=8.577, p=.003**), with post-hoc analysis indicating that ESC was better than the combinations of CON and DQD (t=4.084, p=.018*); INS and EXP (t=3.869, p=.04*), and ESC was better than the combination of ESC, INS, CON, and EXP (t=4.084, p=.018*).

When it comes to value, the results indicate that few single techniques had a significant effect on value, with DQI performing better than when no technique was used (F= 4.194, p=.041*. Figure 3-c) and CON performing worse than baseline (F=4.430, p=.036*, Figure 3-d). Combinations of two techniques showed that the combination of DQD and DQI had a significant effect on value (F=9.437, p=.002**), with DQI performing better than no technique (t=-3.620, p=.002**) and better than the combination of DQD*DQI (t=2.735, p=.033*). The combination of DQD and EXP also showed a significant effect on value (F=5.440, p=.02*), but post-hoc analysis did not reveal any significant differences. Combinations of three, four, and five techniques did not show significant effects on value. Despite significant effects after ANOVA, post-hoc tests did not yield significant differences (p>.05).

We also tested the effects of the techniques on a dataset averaged across raters to account for variance in ratings. This resulted in 3 times fewer observations, leading to a higher threshold to reach significance in pairwise comparisons. Thus, only ESC showed a positive effect on novelty ratings (F=7.588, p=.007**).

We investigated the effect of prompt length on output ratings. Correlation analyses show no difference in ratings based on prompt length (rho=.053, p=.465).

5. Discussion

5.1 Answer to the research question

This study investigated the impact of using various prompting techniques on the novelty and value of ideas generated by GPT-3. The techniques tested included Evaluation Criteria Specification (ESC), Context (CON), Instruction/Schema (INS), Demonstration: Question Dependent (DQD), Demonstration: Question Independent (DQI), and Expert Perspective (EXP). The results showed that ESC had a positive impact on the novelty of ideas, while Context had a negative impact. Our study seems to align with previous findings of Summers-Stay et al. (2023), which suggest that idea quality may be improved by adjusting the prompt. They speculate that such systems may have picked up on creative ideas for other brainstorming questions in the training data and point to the capabilities of such models for analogical reasoning. Our prompt with ESC asking for "novel ideas" might then trigger the GLM to produce such ideas. Given the complexity and opacity of such models, the mechanisms for producing such results, however, remain speculative and require further testing. Combinations of two or more techniques showed significant effects on novelty, with some techniques working better together than others. However, a few single techniques had a significant impact on the value of ideas, and combinations of techniques did not show any significant effects on value. The study's findings suggest that using specific combinations of prompting techniques in idea generation with generative AI can enhance the novelty of ideas generated, but further research is needed to determine their impact on idea value. This further aligns with earlier findings of Summer-Stay et al. (2023), who also only found a small difference through their more advanced approach aiming for increasing utility (i.e., a dimension of *value*). As stated before, given the complexity of such models and the scarcity of empirical research (Liu et al., 2023), one can only speculate as to the reason.

Additionally, the results showed that Evaluation Criteria Specification (ESC) performed better when used alone rather than in combination with other techniques. This finding suggests that adding many techniques to the prompt may not necessarily enhance idea quality (over-engineering). A potential explanation might be "spurious correlations" (Brown et al., 2020, p. 7), a problem that can occur when adding information beyond the task (e.g., examples). Instead, it may be more effective to use a targeted approach and select a few specific techniques based on the desired outcome. This highlights the importance of prompt engineering for creative tasks, which may require a different approach than other use cases where combinations of prompts can improve results (Mishra, Khashabi, Baral, & Hajishirzi, 2022; Wu et al., 2022). Therefore, it is important to carefully consider the use of multiple techniques and assess their impact on idea quality based on the specific context and goals of the task.

Our findings of only a few statistically significant differences regarding the use of prompt engineering techniques are somewhat surprising, contrary to anecdotal evidence from social media stressing the importance of prompt engineering. One reason why we find only a few differences between the prompts could be that we produce relatively short output in a restricted format (i.e., 3 bullet points as opposed to large blocks of free text). We observed many variations of the seemingly same idea. Potentially, there are some "obvious" answers that are provided first, essentially "overshadowing" prompt differences. However, this is speculation, and further research is required, e.g., using *calibration*, as discussed by Liu et al. (2023). Other potential reasons for us observing only a few significant differences are discussed in the limitations.

5.2 Contribution to theory

From a theoretical perspective, the study adds to the growing body of literature on using GLMs productively in creative applications (Gero et al., 2022; Shakeri et al., 2021; Yuan et al., 2022), particularly for the case of generating ideas in brainstorming-like settings (Di Fede et al., 2022; Haase & Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023). Such idea-generation approaches typically rely on a specific, pre-defined prompt template, which may (or may not) include user-generated ideas. With this study, we enhance the understanding of the role of prompt engineering (Mishra, Khashabi, Baral, & Hajishirzi, 2022; Wu et al., 2022). Such an understanding is important, as there is an increasing number of products that embed GLMs. For such systems, a crucial aspect of system design is to formulate an appropriate prompt template. In doing so, system designers essentially help the user to ask the GLM "the right questions" (Dang et al., 2022).

By examining the effect of different single and combined techniques on both idea novelty and value, we enhance the understanding of how to design systems for effective ideation with GLMs. The finding that ESC performed better when used alone rather than in combination with other techniques challenges the assumption that combining many techniques enhances the output quality. It rather highlights the need to consider using multiple techniques carefully in creative tasks and the importance of prompt engineering.

On a more abstract level, we contribute to the field of CSS by investigating the next-generation GLM-driven systems for supporting human creativity. Such a

GLM-based brainstorming system was also suggested by Di Fede et al. (2022); with our analysis, we can inform the underlying technical design.

5.3 Implications for practice

From a practical perspective, the study offers valuable insights for organizations seeking to enhance their innovation potential. By identifying which techniques are most effective in enhancing the idea novelty and value, organizations can design more effective ideation sessions and improve their creative output. The finding that ESC has a positive impact on novelty highlights the importance of clearly defining brainstorming evaluation criteria to shift the focus on specific goals to generate more creative ideas. Additionally, the finding that combining DQI with EXP leads to more novel ideas than CON alone has practical implications for the use of expert perspectives in idea generation. By combining the expertise of internal and external experts with demonstrational prompts, organizations can generate more innovative ideas. However, results are to be interpreted carefully, as our study so far only indicates potential trends, but no general conclusions can be drawn.

There are large differences in the length of the prompts (14 to 614 words), particularly due to the *question-independent demonstrations* or *context* techniques. When costs for GLM usage are based on prompt length, this can result in cost differences for the prompt templates, which should be considered by tool designers, given the small differences in output quality.

5.4 Limitations & outlook

With only a few statistically significant differences between prompt engineering techniques, our results run counter to our (and many anecdotal) expectations. However, the results are to be interpreted carefully due to several limitations of our study that we discuss below.

While we based the candidate prompt development on existing literature, there are many ways of operationalizing the different prompt engineering techniques, and different implementations may result in different outputs. We, however, only tested one operationalization per technique, prioritizing testing more techniques over more operationalizations per technique. Additionally, the field of prompt engineering is rapidly evolving, and new techniques are constantly suggested. Thus, future research may expand on our results for both adjusting the operationalization of techniques or including additional techniques.

We have selected GPT-3 because it is a powerful, widely adopted model (Brown et al., 2020). However, there are other GLMs, and GLMs are constantly evolving, which, in some cases, affects the quality of GLM-generated ideas (Haase & Hanel, 2023). Additionally, there is a dependency between the effectiveness of prompts and model sizes (Liu et al., 2023), which could affect prompt techniques and transferability of results; some even suggest prompt engineering might become obsolete when models improve (Oppenlaender et al., 2023). Thus, future research should explore how different model (versions) together with prompt techniques affect the idea quality.

Having 64 templates, each technique was present in 32 prompts. With 3 ideas per prompt, we had 96 ideas with and without each prompt technique, allowing for meaningful comparisons. However, including combinations of techniques reduced the number of observations per group. Tukey's post-hoc test takes into account the variability of the data and sets a threshold for statistical significance that is adjusted for the number of pairwise comparisons being made, e.g., if there are 16 pairwise comparisons, the adjusted p-value is 0.05/16=0.003, thus making it harder to spot significant differences. To counteract this, future studies could rely on these results to investigate fewer techniques and additionally collect a higher number of observations.

Originally, we had planned to assess the robustness of our results across three societal problem questions; however, due to the low level of agreement among the crowd workers' responses, we had to use a panel of three raters to rate all ideas. To avoid fatigue, ideas for only one scenario were assessed (192 ideas on two dimensions). We thereby prioritized testing technique combinations and high data quality over additional scenarios. Thus, we cannot report insights on the robustness across problem questions. While the nature of the assessed information system – a general-purpose GLM without adaptions to our scenario – conceptually does not give any particular reason to assume that it would function completely differently across scenarios that are relatively common knowledge, future research should investigate these results to increase robustness.

For assessing idea quality, we selected two evaluation criteria (novelty, value) that reflect the criteria commonly used in literature. However, there are other criteria, such as practicability or plausibility of ideas (incl. costs), as well as criteria for sets of ideas such as *diversity* (Siangliulue et al., 2015), or even more question-specific criteria, such as social utility or moral value. Future research should further explore how GLMs could support idea generation along such criteria.

Lastly, inter-rater agreement for the three-rater panel was higher than for crowd-workers. However, an even better agreement, particularly for 'value', might have surfaced more statistically significant findings.

Given these limitations, we call for further investigation of different ways of operationalizing the prompt engineering techniques and an application to other problems to increase robustness. Our approach is transferable to other brainstorming questions and can serve as a foundation. Additionally, we suggest increasing inter-rater agreement, e.g., by increasing the number of raters or by using an evaluation scheme. More broadly, the ideation performance of humans working with such systems should be investigated.

While not at the core of our study, future research needs to investigate the implications of GLMs' inability to *understand* language and develop mitigation strategies for the issues of biases, falsehoods, and lack of moral judgment – discussed more broadly for GLMs (Bender et al., 2021; Floridi & Chiriatti, 2020; Lin et al., 2022; Susarla et al., 2023) – for creativity- and work-related use, particularly as brainstorming sessions are framing problem and solution corridors. Depending on the specific problem at hand, biased GLMs’ suggestions could have severe negative consequences. We thus urge future research to explore sociotechnical perspectives of integrating generative AI in work settings by leveraging human expertise appropriately and preventing humans from accepting unfit suggestions, potentially inspired by existing research on the engagement of humans with AI systems’ outputs from the area of AI-assisted decision-making (e.g., Buçinca et al., 2021).

6. Conclusion

GLM capabilities may be embedded into products, abstracting away the difficulties of interacting with the GLM directly while enabling unskilled users to still benefit from it. While this frees the user from developing a *good* prompt, it makes the design of a good *prompt template* during system development necessary. In our study, we develop a set of 64 prompt templates according to six prompt engineering techniques for an ideation tool, generate ideas via the GLM ‘GPT-3’ for an exemplary societal problem, and have these ideas evaluated. We find that prompt engineering techniques only in a few cases significantly affect idea quality positively. As this is counter to popular belief and our own experience, we carefully discuss potential explanations and limitations of our study. Particularly, we encourage further research on the effects of using prompt design for creative, divergent thinking tasks.

Acknowledgment. This research was funded by the German Federal Ministry of Education and Research (BMBF) in the context of the project HyMeKI (reference number: 01IS20057).

7. References

Althuizen, N., & Reichel, A. (2016). The Effects of IT-Enabled Cognitive Stimulation Tools on Creative Problem Solving: A Dual Pathway to Creativity. *Journal of Management Information Systems*, *33*(1), 11–44. https://doi.org/10.1080/07421222.2016.1172439

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., . . . Amodei, D. (2020). Language Models are Few-Shot Learners. In *NIPS'20: Proceedings of the 34th International Conference on Neural Information Processing Systems* (pp. 1877–1901). https://doi.org/10.5555/3495724.3495883

Buçinca, Z., Malaya, M. B., & Gajos, K. Z. (2021). To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–21. https://doi.org/10.1145/3449287

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment*, *6*(4), 284–290. https://doi.org/10.1037/1040-3590.6.4.284

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022, September 3). *How to Prompt? Opportunities and Challenges of Zero- and Few-Shot Learning for Human-AI Interaction in Creative Applications of Generative Models*. http://arxiv.org/pdf/2209.01390v1

Di Fede, G., Rocchesso, D., Dow, S. P., & Andolina, S. (2022). The Idea Machine: LLM-based Expansion, Rewriting, Combination, and Suggestion of Ideas. In *Creativity and Cognition* (pp. 623–627). ACM. https://doi.org/10.1145/3527927.3535197

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, *30*(4), 681–694.

Frich, J., MacDonald Vermeulen, L., Remy, C., Biskjaer, M. M., & Dalsgaard, P. (2019). Mapping the Landscape of Creativity Support Tools in HCI. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–18). ACM. https://doi.org/10.1145/3290605.3300619

Gabriel, A., Monticolo, D., Camargo, M., & Bourgault, M. (2016). Creativity support systems: A systematic mapping study. *Thinking Skills and Creativity*, *21*, 109–122. https://doi.org/10.1016/j.tsc.2016.05.009

Gero, K. I., Liu, V., & Chilton, L. (2022). Sparks: Inspiration for Science Writing using Language Models. In F. `. Mueller, S. Greuter, R. A. Khot, P. Sweetser, & M. Obrist (Eds.), *Designing Interactive Systems Conference* (pp. 1002–1019). ACM. https://doi.org/10.1145/3532106.3533533

Haase, J., & Hanel, P. H. P. (2023). *Artificial muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity*. https://doi.org/10.48550/arXiv.2303.12003

JASP. (2023). *JASP (Version 0.17)[Computer software]*. https://jasp-stats.org/

Jiang, E., Olson, K., Toh, E., Molina, A., Donsbach, A., Terry, M., & Cai, C. J [Carrie J.] (2022). PromptMaker: Prompt-based Prototyping with Large Language Models. In S. Barbosa, C. Lampe, C. Appert, & D. A. Shamma (Eds.), *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1–8). ACM. https://doi.org/10.1145/3491101.3503564

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

Maaravi, Y., Heller, B., Shoham, Y., Mohar, S., & Deutsch, B. (2021). Ideation in the digital age: literature review and integrative model for electronic brainstorming. *Review of Managerial Science*, *15*(6), 1431–1464. https://doi.org/10.1007/s11846-020-00400-5

Memmert, L., & Tavanapour, N. (2023). Towards Human-AI-Collaboration in Brainstorming: Empirical Insights into the Perception of working with a generative AI. In *31st European Conference on Information Systems.* https://aisel.aisnet.org/ecis2023_rp/429

Mishra, S., Khashabi, D., Baral, C., Choi, Y., & Hajishirzi, H. (2022). Reframing Instructional Prompts to GPTk's Language. In *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 589–612).

Mishra, S., Khashabi, D., Baral, C., & Hajishirzi, H. (2022). Cross-Task Generalization via Natural Language Crowdsourcing Instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics.*

Nelissen, E. R. (2022). *Improving individual idea generation with the selective brainwriting technique.*

Oppenlaender, J., Linder, R., & Silvennoinen, J. (2023). *Prompting AI Art: An Investigation into the Creative Skill of Prompt Engineering.* https://doi.org/10.48550/arXiv.2303.13534

Pilcicki, R., Siemon, D., & Lattemann, C. (2022). How Feature- and Communication Constraints in CSS Affect Creative Collaboration in Virtual Teams - An Activity Theory Perspective. In *55th Hawaii International Conference on System Sciences*.

Przybilla, L., Baar, L., Wiesche, M., & Krcmar, H. (2019). Machines as Teammates in Creative Teams. In D. Joseph, C. van Slyke, J. P. Allen, J. Quesenberry, & M. Wiesche (Eds.), *Proceedings of the 2019 on Computers and People Research Conference* (pp. 94–102). ACM.

Reynolds, L., & McDonell, K. (2021). Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In Y. Kitamura, A. Quigley, K. Isbister, & T. Igarashi (Eds.), *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–7). ACM. https://doi.org/10.1145/3411763.3451760

Schallmo, D., & Lang, K. (2020). *Design Thinking erfolgreich anwenden: So entwickeln Sie in 7 Phasen kundenorientierte Produkte und Dienstleistungen* (2., aktualisierte Auflage). Springer Gabler. https://doi.org/10.1007/978-3-658-28325-4

Shakeri, H., Neustaedter, C., & DiPaola, S. (2021). SAGA: Collaborative Storytelling with GPT-3. In J. Birnholtz, L. Ciolfi, S. Ding, S. Fussell, A. Monroy-Hernández, S. Munson, I. Shklovski, & M. Naaman (Eds.), *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing* (pp. 163–166). ACM. https://doi.org/10.1145/3462204.3481771

Siangliulue, P., Chan, J., Gajos, K. Z., & Dow, S. P. (2015). Providing Timely Examples Improves the Quantity and Quality of Generated Ideas. In T. Maver & E. Y.-L. Do (Eds.), *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition* (pp. 83–92). ACM.

Stevenson, C., Smal, I., Baas, M., Grasman, R., & van der Maas, H. (2022, June 10). *Putting GPT-3's Creativity to the (Alternative Uses) Test.*

Summers-Stay, D., Voss, C. R., & Lukin, S. M. (2023). Brainstorm, then Select: a Generative Language Model Improves Its Creativity Score. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus Effect of Generative AI: Charting the Path for Responsible Conduct of Scholarly Activities in Information Systems. *Information Systems Research*, *34*(2), 399–408. https://doi.org/10.1287/isre.2023.ed.v34.n2

Wu, T., Terry, M., & Cai, C. J [Carrie Jun] (2022). AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In S. Barbosa, C. Lampe, C. Appert, D. A. Shamma, S. Drucker, J. Williamson, & K. Yatani (Eds.), *CHI Conference on Human Factors in Computing Systems* (pp. 1–22). ACM. https://doi.org/10.1145/3491102.3517582

Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022). Wordcraft: Story Writing With Large Language Models. In *27th International Conference on Intelligent User Interfaces* (pp. 841–852). ACM. https://doi.org/10.1145/3490099.3511105

Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021). Calibrate Before Use: Improving Few-shot Performance of Language Models. In *Proceedings of the 38th International Conference on Machine Learning*.

Zhu, Q., & Luo, J. (2022). Generative Pre-Trained Transformer for Design Concept Generation: An Exploration. *Proceedings of the Design Society*, *2*, 1825–1834. https://doi.org/10.1017/pds.2022.185

Zhu, Y., Ritter, S. M., & Dijksterhuis, A. (2020). Creativity: Intrapersonal and Interpersonal Selection of Creative Ideas. *The Journal of Creative Behavior*, *54*(3), 626–635. https://doi.org/10.1002/jocb.397

Zhu, Y., Ritter, S. M., & Dijksterhuis, A. (2021). The effect of rank-ordering strategy on creative idea selection performance. *European Journal of Social Psychology*, *51*(2), 360–376. https://doi.org/10.1002/ejsp.2743