Human-AI Collaboration for Brainstorming: Effect of the Presence of AI Ideas on Breadth of Exploration

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

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

Abstract

With the widespread adoption of generative large language models (GLMs) such as GPT-3 or ChatGPT for human-AI problem solving, understanding the effect on performance becomes important. Brainstorming is an established approach for generating ideas to solve problems. In this study, we investigate how AI ideas affect the brainstorming performance metric 'flexibility', which refers to the breadth of exploration or coverage of the topic. The foundation for our analysis is the data from an experiment (n=52) in which individual participants brainstormed in two conditions: (1) human-only (baseline) and (2) human+AI (treatment). The treatment condition had access to ideas generated via the GLM OpenAI GPT-3.5. Results show significantly higher flexibility for the human+AI as compared to the human-only condition with a large effect size. With our study, we contribute to the literature of electronic brainstorming, brainstorming with GLMs, as well as to the research challenge of human-AI collaboration.

Keywords: Brainstorming, Human-AI Collaboration, GPT-3.5, Generative Language Models, Performance

1. Introduction

Brainstorming is a popular technique for groups of humans to generate ideas for solving problems (Osborn, 1953). However, other humans are not always cost-effectively available. While using technology to facilitate human brainstorming groups is common, tools to support humans with actual ideas (e.g., Siemon et al., 2015) are scarce. With recent advancements of generative large language models (GLMs), however, new opportunities might arise. GLMs were used for creative tasks (Gero et al., 2022), both in free form, such as in the use of ChatGPT, as well as in embedded form with a use case specific graphical user interface. More specifically, several studies successfully used GLMs such as ChatGPT, GPT-3, or GPT-4 for generating ideas (Haase & Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023), even suggesting that GLMs' creative abilities might be comparable to human creative abilities (Haase & Hanel, 2023). Thus, in this study, we go beyond considering the technical system merely as a tool as well as the human and the technical system separately but instead explore human-AI collaboration. In an approach similar to Di Fede et al. (2022), we set up a GLM-based brainstorming app, allowing users to request idea suggestions from a GLM for the brainstorming question. The GLM might thus benefit the human by contributing ideas similar to a human (Haase & Hanel, 2023) or by providing inspiration.

However, such GLMs have limitations. With many of them being trained on large datasets from the internet, GLMs may reproduce bias from the training data and show limited output diversity (Bender et al., 2021; Lin et al., 2022). This calls into question how effectively such systems can be used in a creative brainstorming setting, for which one outcome measure frequently used is flexibility, i.e., breadth of exploration or coverage of a topic (Althuizen & Reichel, 2016; Nijstad et al., 2010). We therefore seek to answer the following research question (RQ): *How does generating ideas jointly with a generative language model affect the breadth of exploration (flexibility)?*

To answer this research question, we used data collected for a previous study (under review) focusing on human-AI brainstorming. We developed a prototype for a GLM-based brainstorming app (Figure 1). We used the powerful OpenAI GPT-3.5 model (OpenAI) as a technical foundation. We had participants brainstorm for 10 minutes on a societal problem. Participants were assigned to either the human-only (baseline) or the human+AI (treatment) condition, enabling a quantitative group comparison. While the previous study focused on comparing the outcomes regarding quantity, novelty, and value of individual ideas, the study at hand focuses on the breadth of exploration and, thereby, on assessing the quality of sets of ideas.

To measure flexibility, we followed the well-established approach of classifying ideas into (pre-defined) categories (Althuizen & Reichel, 2016; Nijstad

URI: https://hdl.handle.net/10125/106425
978-0-9981331-7-1

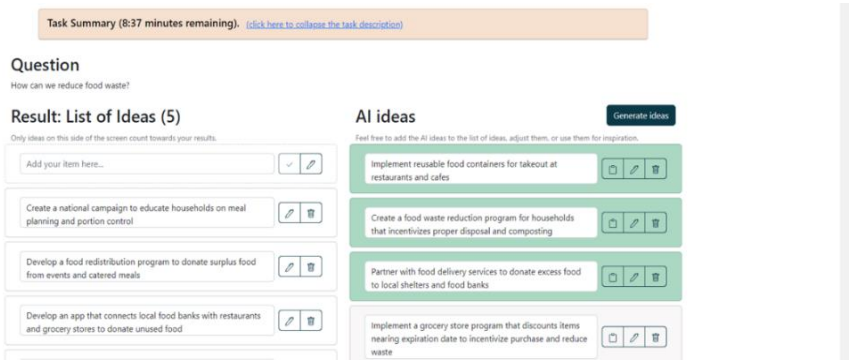

Figure 1. GLM-based brainstorming app prototype for data collection

et al., 2010; Ritter & Mostert, 2018). Counting the number of categories covered by the ideas generated by the participants provides a measure of how broadly the problem was addressed. We statistically performed a group comparison on the breadth of coverage depending on the condition. Additionally, we perform a more in-depth analysis on the breadth of ideas proposed by the AI and assess who drives the exploration of new categories. We find the human+AI team covers the brainstorming question significantly more broadly as compared to the humans who work alone.

With our work, we contribute to the long research history of brainstorming research (Osborn, 1953), more specifically, using technology to enhance brainstorming performance (Pinsonneault et al., 1999). With our tool, we move beyond merely facilitating brainstorming (on a meta-level) but examine a system that supports through actual ideas similar to a human (content-level).

Additionally, we contribute to the emergent literature of exploring challenges, limitations, or boundaries when using GLMs (Floridi & Chiriatti, 2020; Lin et al., 2022). We find that while GLMs were shown to reproduce bias, the output on our societal problem was diverse with regard to the categories covered. We thereby sharpen the understanding of how GLMs may support work-related idea generation.

More broadly, we contribute to the emergent challenge of problem-solving through human-AI collaboration (Akata et al., 2020; Dellermann et al., 2019; Krogh, 2018) and improving the understanding of team dynamics of humans working alongside AI (Makarius et al., 2020), showing how turning individual work of a human into collaborative work of human and AI affects an important brainstorming metric.

2. Background

2.1 Brainstorming

Brainstorming is a popular approach for groups of humans to generate ideas for solving problems (Osborn, 1953). For effective brainstorming, Osborn (1953) suggested four rules to follow: (1) delayed judgment, (2) encouragement of wild ideas, (3) quantity of ideas, and (4) combining and improving each other's ideas. Since then, lots of research focused on understanding and improving brainstorming performance.

Brainstorming performance can be measured in a variety of ways. Typical measures include the quantity ("fluency"; Nijstad et al., 2010) and quality of ideas produced. Quality can be assessed on the level of individual ideas, e.g., considering the novelty or value of ideas (previous study). Additionally, quality can be assessed on the level of sets of ideas, which we will focus on in this study. An important measure is how broadly the brainstorming question is covered (Althuizen & Reichel, 2016; Nijstad et al., 2010). A typical way to conceptualize coverage is by imagining there to be different aspects or categories of ideas for a brainstorming question. A typical way for assessing the breadth of coverage is by explicitly developing a category system for ideas to reflect the different aspects of a topic and sorting all ideas of the brainstorming session into these developed categories. For each brainstorming session, the breadth can then be calculated as the number of distinct categories covered in this session, from zero (no idea) to the number of categories within the category system. The category system can be developed inductively based on the ideas (e.g., Althuizen & Reichel, 2016) or adapted from prior research.

Past decades of research surfaced many performance-enhancing and -reducing brainstorming effects (Pinsonneault et al., 1999). While Pinsonneault et al. (1999) list 16, we here describe the two most relevant effects for our research question: cognitive stimulation (performance-enhancing) and cognitive inertia (performance-reducing).

Cognitive stimulation refers to the effect that the "utterance of [brainstorming group] members may contain task related stimuli that elicit new ideas from other members" (Pinsonneault et al., 1999). Providing

stimuli, e.g., through a confederate or played via audio tape, can improve brainstorming performance (Dugosh et al., 2000; Paulus et al., 2013). However, cognitive stimulation only occurs if humans pay attention to the stimuli (Leggett Dugosh & Paulus, 2005). Additionally, stimuli can have different effects on the way the problem or solution space is explored, with conceptually more distant stimuli encouraging a broader exploration, whereas more closely related stimuli encourage a more in-depth exploration (Althuizen & Reichel, 2016; Althuizen & Wierenga, 2014).

Cognitive inertia, on the other hand, refers to group members "embark[ing] on a single train of thought, which limits creativity and productivity" (Pinsonneault et al., 1999). Typically, cognitive inertia is thought to occur in *nominal groups*, i.e., when group members brainstorm individually, and their ideas are pooled afterward (as compared to brainstorming together as part of *real groups*). In nominal settings, group members do not benefit from outside stimuli and might produce more ideas that are similar, i.e., exploiting existing categories instead of discovering new categories, resulting in a higher (average) number of ideas in only a few categories (higher *within-category fluency*; Nijstad et al., 2010). We discuss both effects with regard to our experiment in the next section.

2.2 Human-AI brainstorming

Besides using AI for decision-making, Krogh (2018) suggests exploring using AI for solving problems. Indeed, more recently, combing humans and AI systems to solve problems was discussed as a new research challenge (Akata et al., 2020; Dellermann et al., 2019). There is a long research history of trying to make brainstorming more effective using technology – researched under the label of electronic brainstorming (Pinsonneault et al., 1999). Such technological support can take the form of facilitation, e.g., through processual guidance or through appropriate design decisions to encourage desired behavior.

Besides such meta-level support, tools contributing on content-level have been explored. Such tools can provide stimuli like words, or partial or full ideas related to the brainstorming question. Such systems might use existing content from social media (Siemon et al., 2015) or curated association dictionaries (Althuizen & Reichel, 2016).

Using AI, however, might be difficult, as beforehand, it is unclear which brainstorming questions the user will want to use the tool for, making gathering training data and training an AI difficult. However, with advances of AI, particularly GLMs, new opportunities arise. GLMs such as GPT-3, GPT-4, or ChatGPT (OpenAI) are trained on a large corpus of text to complete the next word given a certain input. Such systems have shown remarkable results on traditional natural language processing tasks (Brown et al., 2020) as well as on creative tasks (Gero et al., 2022; Q. Zhu & Luo, 2022). Such systems are pre-trained, requiring no task-specific training or fine-tuning (Brown et al., 2020). More specifically, their creative potential was explored for brainstorming (Haase & Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023).

Going beyond investigating the technical system and the human separately, Di Fede et al. (2022) proposed to leverage GPT-3 to support humans in brainstorming. After demonstrating the feasibility of using GPT-3 in such human-AI brainstorming settings (Memmert & Tavanapour, 2023), we now quantitatively investigate this setting.

One important goal for human-AI collaboration is the *superior performance* of the human-AI team as compared to the individual (Dellermann et al., 2019). We thus propose to investigate how turning individual work of humans into collaborative work of humans and AI systems affects performance. Besides the direct effect of the GLM proposing ideas, adding a GLM might also affect the human's idea generation. Earlier, the two effects (*cognitive stimulation* and *cognitive inertia*) were introduced. These effects are known from all-human groups. Given that humans are known to respond socially to technical systems (Nass & Moon, 2000) and GLMs' creative abilities were described to be comparable to humans' creative abilities (Haase & Hanel, 2023), we suggest applying this lens to our collaborative human-AI setting. However, it is unclear if those effects occur in such an interactive human-AI setting and how they affect performance overall.

As part of this study, we provide participants access to ideas for a brainstorming question, which are generated by a GLM. Such suggestions could act as stimuli, which might lead to cognitive stimulation. In a previous study (Memmert & Tavanapour, 2023), participants reported having felt inspired to explore new areas of the problem (i.e., *cognitive stimulation*). This might lead to the assumption that access to such AI suggestions could increase flexibility. However, other participants reported that the AI suggestions were basically their previous ideas rephrased in other words. In this case, the AI ideas (i.e., stimuli) might not be capable of preventing *cognitive inertia*, typically observed when working alone. On the contrary, showing potential examples of solutions might lead to *fixation* (Lamm & Trommsdorff, 1973; Sio et al., 2015).

As discussed before, the conceptual distance of stimuli affects how broadly or deeply the problem and solution spaces are explored (Althuizen & Reichel, 2016). The nature of GLM's suggestions with regard to conceptual distance, however, is unclear. Perhaps more

importantly, there are discussions around GLMs' output quality, e.g., GLMs were shown to reproduce biases, stereotypes, and falsehoods, with limited diversity in outputs (Bender et al., 2021; Floridi & Chiriatti, 2020; Lin et al., 2022). Particularly, the lack of diversity (Bender et al., 2021) might lead to a narrow exploration, which could result in reduced brainstorming flexibility.

Following the brainstorming rules, in our experiment, the ideas of the human are provided as input to the AI, and humans will have access to the AI ideas. As discussed, the implications on flexibility are unclear, both on the level of the human and on the team overall (human+AI). In previous research on brainstorming with GPT-3, participants subjectively reported signs of both a broadening and narrowing of the perspective due to the AI (Memmert & Tavanapour, 2023); additionally, the conceptual distance and diversity of GLM outputs are unclear. We thus propose the following two undirected hypotheses:

H1. Presence of AI support will affect the diversity of ideas on team level.

H2. Presence of AI support will affect the diversity of ideas on individual level.

While many comparisons might be feasible, we compare performance depending on the presence of AI ideas, resulting in the two conditions: human-only vs. human+AI. A human-AI team achieving superior performance as compared to the individual human is a core ambition of human-AI collaboration research (Dellermann et al., 2019). Additionally, from a practical perspective, other humans might not always be cost-effectively available, which is why we did not include a comparison to human teams at this stage.

3. Method

3.1 Procedure and participants

For data collection, we developed a GPT-based brainstorming app (see Figure 1 for a screenshot and the next subsection for a more detailed description), similar to the app proposed by Di Fede et al. (2022). We had participants brainstorm individually for 10 minutes on a societal problem. Societal problems are common for such brainstorming studies (Huber et al., 2019). We selected the problem of food waste reduction ('How can we reduce food waste?'), which was used in previous brainstorming studies (Y. Zhu et al., 2021). Afterward, we had participants select their best ideas and answer a post-questionnaire. Participating in the study took about 30 minutes. We conducted the study in three courses taught at our university's informatics department on (1) design science research (two sessions), (2) computer-supported cooperative work, and (3) data-driven solutions for the smart city.

For our experiment, we had two conditions: a baseline condition (human-only), in which each human brainstormed individually, and a treatment condition (human+AI), in which each human brainstormed together with a GLM. As we sought to understand how the GLM affects the results, we asked participants to request suggestions at least once. We communicated to the participants that only suggestions in the "list of ideas" (left side of the screen) would count toward their results (i.e., not-accepted AI ideas are excluded).

3.2 Data collection instrument

We collected the data for a different study (under review) with a focus on understanding how AI suggestions affect idea quantity and quality. To collect the data, we developed a brainstorming app containing the study details, the brainstorming question, and a timer. Participants could add, edit, and remove ideas. Participants in the human+AI condition could request AI suggestions, which were displayed next to their own ideas in a separate list (3 ideas at a time). We decided to allow copying AI suggestions into the participants' "list of ideas" instead of only using suggestions as stimuli to make the scenario more realistic. However, the GLM might also fulfill the role of a stimuli provider.

```
You are part of a brainstorming team. Your
goal is to come up with novel and valuable
ideas for the following question:
{BRAINSTORMING_QUESTION}
Discarded ideas so far:
{list_ai_ideas}
Novel and valuable ideas so far:
{long_list_ideas}
Please come up with {NUMBER_OF_IDEAS}
additional novel and valuable ideas for
the question: "{BRAINSTORMING_QUESTION}".
Please provide the {NUMBER_OF_IDEAS}
additional ideas as enumerated, ordered
list. Each idea should be limited to a
maximum of 20 words.
```

Figure 2. Prompt template

Note: curly brackets indicate placeholders; all caps placeholder labels indicate parameters fixed for this study; small caps placeholder labels indicate parameters changing based on the current canvas state; portions in italic font are only included if ideas are already present in the respective lists

To produce the AI suggestions, we used OpenAI's GLM solution 'gpt-3.5-turbo-0301' at a temperature of 0.9, as recommended for creative applications (OpenAI Documentation). We defined a prompt template (see Figure 2) and dynamically populated it with the study details (i.e., brainstorming question) and the current state of the brainstorming pane, i.e., content entered by the participant (ideas) and past AI ideas. The human ideas are provided to the GLM, as an important aspect

of brainstorming in groups is to build on each other's ideas (Osborn, 1953). Such reciprocity is also a core aspect of collaboration (Bedwell et al., 2012).

We decided to make users actively request suggestions instead of pushing suggestions proactively, as this was shown to be most effective (Siangliulue et al., 2015). All tool interactions are logged for analysis.

3.3 Data preparation and analysis

To investigate the diversity of ideas, we followed the established approach of sorting all ideas into categories (Althuizen & Reichel, 2016; Nijstad et al., 2010; Ritter & Mostert, 2018). After reviewing multiple category systems on 'food waste', we decided to use the system of Specht and Buck (2019). The system consists of 13 categories within five clusters. We selected this category system as it was developed on user-generated contributions (tweets), which are similar in length to the ideas in our study, based on contributions from users from a Western country (US; our participants attend a Western European university), and had an appropriate granularity, i.e., number of categories (compare, e.g., Althuizen & Reichel, 2016). We had a blind-to-condition student assistant sort the ideas into the pre-defined categories (see Table 1). All but seven ideas were categorized accordingly. We excluded all non-categorized and all deleted ideas from the analysis.

For data analysis, we processed the log data with Python scripts. To calculate descriptive and inferential statistics (incl. assumption checks), we used JASP (JASP, 2023). For visualizations, we used Tableau.

4. Results

Our study had 54 participants. We had to exclude the data of two students (due to incomplete data and due to a misunderstanding with regard to the task). The remaining 52 participants (age: mean=22.3, min=18, max=37 years old; gender: 8 female, 44 male) were university students enrolled mostly in informatics study programs. Participants were randomly assigned to conditions and were equally distributed across conditions (human-only: 26, human+ai: 26). In total, 600 ideas were included in the analysis. The distribution across conditions and idea origin is shown in Table 2. Participants in the human+ai condition requested AI input 148 times (mean=5.7), but only a subset of these ideas is reflected in the final idea set.

Table 2. Number of ideas (and mean values) by origin and condition

Origin Condition	AI	human	Total
human+AI	373 (mean=14.3)	87 (mean=3.3)	**460**
human-only	-	140 (mean=5.4)	**140**
Total	**373**	**227**	**600**

4.1 Breadth of exploration

We investigate our first hypothesis, exploring whether 'superior performance', an important aspect of human-AI collaboration, was achieved in our setting. Superior performance here refers to whether the human-AI team performed better than the human individually, with respect to the metric 'flexibility'.

We calculate the breadth of coverage for each participant (including the 115 AI suggestions that were

Table 1. System of categories (Specht & Buck, 2019) and exemplary ideas from brainstorming sessions

Area	Category	Exemplary ideas from brainstorming sessions
Domestic or household behavior change	Meal planning	Promote meal planning and portion control to reduce overbuying and food waste at home
	Waste mitigation	Consuming leftover products instead of buying new products
	Smart technology	Utilize technology to create smart refrigerators that track food expiration dates and provide recipe suggestions using expiring ingredients
Food waste diversion and donation	Large-scale food donation	Create a food donation program for excess food from commercial kitchens and events
	Food waste markets	Implement "ugly produce" programs that sell visually imperfect but still edible fruits and vegetables at a discount
Recycling and upcycling	Value-added products	Develop biodegradable packaging made from food waste materials
	Converting food waste into energy	Create a network to distribute food waste from grocery stores and restaurants to biogas facilities for renewable energy production
	Food waste for agricultural purposes	Utilize food surplus for sustainable animal feed to reduce waste in the agriculture industry
Consumer education	Public information campaigns	Create a national awareness campaign that educates consumers on the environmental and social impact of food waste
	Mobile technology	Develop an app that connects individuals with nearby food businesses to donate excess food before it spoils
	Family and consumer science training	Create a food preservation education program to teach individuals how to properly store and preserve food
Governmental action and policy	Legislating food waste reduction	Encourage food donation by providing tax incentives for businesses that donate excess food
	Food date labeling	Implement a food packaging labeling system that indicates the actual shelf life instead of a standardized date

accepted/copied by the participant for the human+ai condition into their "list of ideas") by counting how many categories of the 13 categories (i.e., distinct) each participant covered (mean=4.115). We then perform a group comparison (see Figure 3). As the normality assumption was not fulfilled for the human-only condition (Shapiro-Wilk: W=.874, p=.004; significant results suggest a deviation from normality), we conducted the non-parametric alternative to the Student's t-test, the Mann-Whitney U test. The test showed a significant difference (U=546.5, p<.001) between the human+AI condition (mdn=5) and the human-only condition (mdn=2), with a strong effect size (rank-biserial correlation r_B=.617). Thereby, the first hypothesis is confirmed, and the important goal of 'superior performance' of the human-AI team is met.

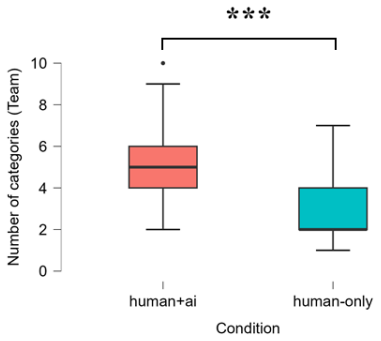

Figure 3. Group comparison on teams' number of categories covered (* p < .001)**

We then investigated our second hypothesis, i.e., that humans themselves (i.e., excluding ideas of AI origin in the human+ai condition) cover the question more broadly as compared to the humans working alone (see Figure 4). For each participant, we calculated the breadth of coverage (excluding AI ideas for the human+ai condition) for the 13 categories (mean=2.846). As the assumption of normality was fulfilled for neither of the conditions (Shapiro-Wilk: human-only: W=.874, p=.004, human+AI: W=.902, p=.018), we calculated the non-parametric alternative to the Student's t-test, the Mann-Whitney U test. We found no significant difference between the two conditions (U=326.0, p=.830). We thus reject our second hypothesis; we do not observe a difference in flexibility for the human with AI ideas present.

A potential explanation could be the number of ideas contributed by the humans. In our previous study (under review), we found humans in the human-only condition to contribute significantly more ideas as compared to the humans in the human+AI condition. Assuming that there might be a tendency to explore more categories the more ideas are contributed, we examined whether a correlation was present. As the normality assumption did not hold, we calculated Spearmen's rho. We find a strong, significant, positive correlation (Spearman's ρ=.817, p<.001), confirming earlier findings of Althuizen and Reichel (2016). Thus, fewer categories explored by the humans in the human+AI condition might be partially attributed to humans contributing fewer ideas in this condition.

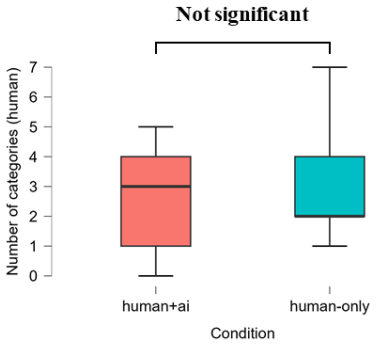

Figure 4. Group comparison on humans' number of categories covered

Humans in the human-only condition contributed significantly more ideas compared to humans in the human+AI condition. However, they did not differ significantly in the number of categories covered. Thus, we expected that the humans in the human-only condition contributed more ideas per category on average (i.e., higher within-category fluency). To test the assumption, we calculated the average number of ideas per category covered by each participant (for participants with no ideas we set 0). The Mann-Whitney U test (assumptions of normality not fulfilled; Shapiro-Wilk: human-only: W=.771, p<.001, human+AI: W=.831, p<.001) shows a significant difference (U=119.0, p<.001) in the average number of ideas per category between human-only (mdn=2) and human+ai (mdn=1) condition. Thus, humans in the human-only condition show higher within-category fluency.

4.2 Driving category exploration

For more context, we performed an analysis on who drove category exploration. To do so, we analyzed, on a summative level, how broadly the categories were explored by both humans and the AI (including 373 AI suggestions). Additionally, we examined who drove the exploration of new categories in the sessions.

For the first analysis, we calculated both for the ideas originating from humans and from the AI how these ideas were distributed across the categories (see Figure 5). As the number of total ideas differ, we show the relative distribution. We find that the AI covers all 13 categories, whereas the humans cover only 12 categories, not providing ideas for the category of 'value-added products'. Across conditions, most ideas

fell into the 'waste mitigation' category, whereas fewest fell into the 'value-added products' category.

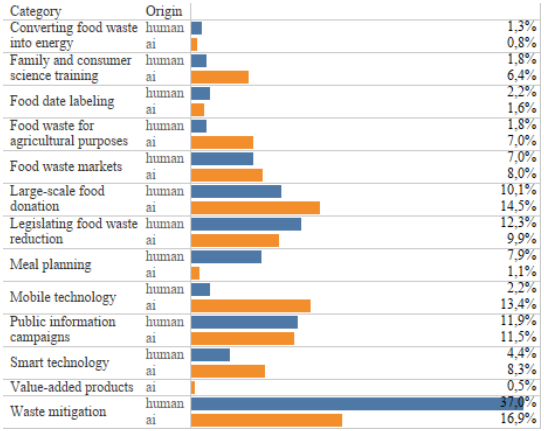

Figure 5. Percentage of ideas across categories of humans (blue) and AI (orange)

On a category-level, we observe the largest difference between human and AI for 'waste mitigation'. Whereas 37.0% of human ideas fell into this category, it was only 16.9% of AI ideas. The AI put a relatively large emphasis (13.4%) on 'mobile technology' compared to the humans (2.2 %). However, these values are not independent, i.e., the humans and the AI system had access to each other's ideas, as suggested per the brainstorming rules (Osborn, 1953).

We thus additionally take a process perspective, asking who proposes ideas of new categories when the brainstorming counterpart has already contributed ideas. This analysis can only be carried out for the human+AI condition. For the analysis, we order the contributions of both humans and AI systems chronologically within each brainstorming session (per participant). We then examined all cases where the human contributed after the AI made a contribution and vice versa. We find that the AI explores new categories in 41.2% of the cases (i.e., when proposing a new idea), whereas the human explores a new category in 41.9% of the cases.

4.3 Subjective perception

To get a more complete understanding of the human+AI collaboration, we asked participants open-ended questions about their experiences. Several participants reported to have felt that the AI helped them explore the topic more quickly and comprehensively:

- "AI can help you to bring more ideas to light" (P29)
- "I was working much faster. Thanks to the AI more ideas came to my mind in less time which made the Brainstorming process much easier than usually" (P28)

Additionally, some participants reported that the AI suggestions helped them to get a new perspective topic, hinting at cognitive stimulation to have occurred:

- "Because of the AI I first came up with certain ideas" (P29)
- "The new ideas of the AI helped me thinking of additional ideas." (P27)
- "The ideas on the side which came from the AI were a great inspiration for developing new ideas on my own." (P28)

However, participants also reported that the AI influenced and potentially narrowed their way of thinking, particularly due to repetitive suggestions, which could hint at cognitive inertia to have occurred:

- "[The AI] would often propose very similar ideas" (P37)
- "It felt a lot faster, but a bit repetitive also since the AI started generating ideas similar to the previous ones" and "[…] It definitely felt like my thought process was being governed by the AI […]" (P52)
- "The AI stopped coming up with original ideas after a few were generated and I couldn't concentrate on coming up with my own ideas" (P34)
- "[…] set my focus in the direction of the suggestions" (P42)

Overall, the feedback fits the quantitative team-level performance results, with many participants stating that the AI helped them to cover the topic more comprehensively. For individual performance, while some participants reported having gained a new perspective, others responded negatively to the AI, with one stating "Maybe [it] cut[…] off my creativity" (P33).

5. Discussion

5.1 Answer to the research question

Flexibility, i.e., breadth of exploration or coverage, is an important brainstorming performance measure, as it is strongly correlated with a high number of high-quality ideas (Nijstad et al., 2010). With the increasing adoption of GLMs (e.g., GPT-3, ChatGPT) for creative tasks in general and generating ideas more specifically, it becomes important to understand how GLMs affect how broadly or narrowly the problem and solution space is explored. Particularly so, as creativity is typically considered a human strength (Dellermann et al., 2019), and AI systems are known to potentially lack diversity in outputs (Bender et al., 2021).

For our setting, we find that the human working with the AI jointly produced significantly more ideas as compared to alone. Both human and AI seem to have

similarly driven the exploration of new categories. This is interesting, given the assumed superiority of humans over AI in creative tasks, but is in line with more recent findings of Haase and Hanel (2023), who found GLMs to have "comparable to human creative abilities" in certain aspects. Furthermore, the AI produced ideas more evenly across the categories. Thus, the brainstorming performance concerning flexibility seems to be affected positively by the presence of the AI.

However, adding the AI did not lead to more ideas on the level of the human. Other than one might expect according to *cognitive stimulation*, humans did not produce more ideas or cover more categories with their ideas. For the former, one reason could be that the participants working with the AI spend time reviewing and selecting ideas, reducing the time available for brainstorming. Such a phenomenon was already described by Pinsonneault et al. (1999, p. 126) and coined the "distraction effect", explained as "individuals […] spending too much time reading others' ideas rather than thinking about new ideas, thus inadvertently limiting their productivity". This could be a potential (partial) explanation as to why no significant difference was observed with regard to the number of ideas for the humans between the conditions. Given that the number of ideas is correlated with the number of categories covered, this might then (partially) explain the lack of a significant difference in the categories covered. However, future research is required on this aspect.

5.2 Contribution and implications

Our work provides theoretical contributions to the literature streams of brainstorming with GLMs, brainstorming group effects as well as human-AI collaboration more broadly and offers practical contributions for supporting brainstorming with GLMs.

We contribute to the discourse around GLMs and brainstorming (Haase & Hanel, 2023; Koivisto & Grassini, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023), enhancing the understanding of the potential of using GLMs for creative idea generation. Other than previous work, we do not investigate humans and GLMs separately but offer insights into joined human-AI brainstorming sessions as suggested, e.g., by Di Fede et al. (2022). Our analysis aligns with and expands on the findings of Haase and Hanel (2023). Not only do GLMs produce ideas of comparable novelty (Haase & Hanel, 2023), but we show that GLMs also drive category exploration comparable to humans.

With our work, we contribute to the literature of electronic brainstorming and related group effects (Pinsonneault et al., 1999). Due to moving beyond an isolated setting in which humans and GLMs work separately, we offer insights into the applicability of known group effects (cognitive stimulation, cognitive inertia) from all-human groups to human-AI groups. We find that participants' reports contain signs of cognitive stimulation, with some stating they explored new areas due to the AI. Thus, GLMs might be capable of taking the role of a stimuli provider. However, we did not observe a broader exploration of the topic by the human. We offer a potential explanation, the distraction effect, which could offset the performance-enhancing cognitive stimulation effect, but future research should explore how this perceived stimulation materializes.

Participant's reports also contained signs of cognitive inertia, with participants explaining that the AI kept them on their train of thought, which could mean the AI-induced fixation (see Lamm & Trommsdorff, 1973). Quantitatively, the AI does not seem to affect flexibility of the individual humans when measured by the breadth of coverage, i.e., the number of categories covered. However, we did observe participants in the human-only condition contributing more ideas within the categories covered (higher within-category fluency), which could hint at them not leaving their "train of thought", i.e., cognitive inertia (Lamm & Trommsdorff, 1973; Pinsonneault et al., 1999). A potential interpretation could be that the GLM led humans to not focus on only a few categories but to add ideas within other (AI-explored) categories. We thus call for more in-depth research to reconcile these observations.

More broadly, in taking a collaborative perspective, we contribute to the research challenge of solving problems through human-AI collaboration (Akata et al., 2020; Dellermann et al., 2019). We demonstrate a core ambition of human-AI collaboration (superior performance) regarding a key brainstorming metric (flexibility). We show the importance of the GLM in driving this performance, which is remarkable, given the traditionally assumed human superiority in creative tasks (Dellermann et al., 2019).

We also offer a practical contribution to brainstorming by describing an instantiated information system supporting humans to develop ideas for a problem. We demonstrate that even without further training data or fine-tuning and without relying on copying existing content from other platforms or on the manual preparation of stimuli (Althuizen & Reichel, 2016; Siemon et al., 2015), modern GLMs can enhance the creative ability of humans when forming a human-AI team. Given our results, one might encourage using GLMs for brainstorming (particularly in the early phases), as this enables humans to more broadly explore the question, which is known to be correlated with a larger number of high-quality ideas. Future research should explore individual differences (e.g., regarding creative ability or topic knowledge).

5.3 Limitations

Our study has several limitations. We only covered one brainstorming question (i.e., food waste), limiting generalizability. However, using societal problems for brainstorming studies is common, and the specific question was used in brainstorming research before (Y. Zhu et al., 2021). We did not adjust the AI system regarding this specific brainstorming question; on the contrary, the question can simply be replaced during the prompt template population (see Figure 2).

The results of our study are dependent on how flexibility is measured. We selected a pre-defined, published category system to increase objectivity. We reported our reasoning for selecting this category system. However, other category systems exist, which might have led to different results. Closely connected to this point: while we did cover idea flexibility, we did not investigate the cultural diversity of suggestions. Some of the ideas might not be representative of all cultures. The category system was based on the tweets from the users from the US; students in our study attended a Western European university). Besides, using a category system is only one approach to assess flexibility; other approaches include, e.g., semantic distances. However, the approach of using a category system is common. Additionally, besides assessing the idea quality on the level of sets of ideas, assessing the effect of using GLMs on the quality of individual ideas is an important area for future research.

On a more technical note, we only used one prompt template. While we used prompt engineering, a different prompt might have yielded different results. The results might be considered a baseline, as we neither optimized for breadth nor depth of exploration. Lastly, we only used one GLM. However, the model is an advancement on a model that showed high performance across several typical and non-typical natural language model tasks (Brown et al., 2020) and the underlying model for the widely adopted OpenAI product ChatGPT (OpenAI), making it highly relevant to practice.

6. Conclusion

GLMs become more widely adopted in work settings to solve problems. When leveraging such systems, it is important to understand the implications on the work results. In our study, we show that teams of humans and AI can outperform humans brainstorming individually on the metric of flexibility, i.e., the breadth of exploration or number of idea categories covered. We contribute to literature on electronic brainstorming (e.g., Althuizen & Reichel, 2016; Pinsonneault et al., 1999), GLMs for brainstorming (Di Fede et al., 2022; Haase & Hanel, 2023; Stevenson et al., 2022; Summers-Stay et al., 2023), and the research challenge of human-AI collaboration for problem-solving (Akata et al., 2020; Dellermann et al., 2019; Makarius et al., 2020).

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

Akata, Z., Balliet, D., Rijke, M. de, Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L., van Harmelen, F., . . . Welling, M. (2020). A Research Agenda for Hybrid Intelligence: Augmenting Human Intellect With Collaborative, Adaptive, Responsible, and Explainable Artificial Intelligence. *Computer*, *53*(8), 18–28. https://doi.org/10.1109/MC.2020.2996587

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

Althuizen, N., & Wierenga, B. (2014). Supporting Creative Problem Solving with a Case-Based Reasoning System. *Journal of Management Information Systems*, *31*(1), 309–340.

Bedwell, W. L., Wildman, J. L., DiazGranados, D., Salazar, M., Kramer, W. S., & Salas, E. (2012). Collaboration at work: An integrative multilevel conceptualization. *Human Resource Management Review*, *22*(2), 128–145. https://doi.org/10.1016/j.hrmr.2011.11.007

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

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid Intelligence. *Business & Information Systems Engineering*, *61*(5), 637–643. https://doi.org/10.1007/s12599-019-00595-2

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

Dugosh, K. L., Paulus, P. B [P. B.], Roland, E. J., & Yang, H. C. (2000). Cognitive stimulation in brainstorming. *Journal of Personality and Social*

Psychology, *79*(5), 722–735. https://doi.org/10.1037/0022-3514.79.5.722

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

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

Huber, B., Shieber, S., & Gajos, K. Z. (2019). Automatically Analyzing Brainstorming Language Behavior with Meeter. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–17.

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

Koivisto, M., & Grassini, S. (2023). Best humans still outperform artificial intelligence in a creative divergent thinking task. *Scientific Reports*, *13*(1), 13601. https://doi.org/10.1038/s41598-023-40858-3

Krogh, G. von (2018). Artificial Intelligence in Organizations: New Opportunities for Phenomenon-Based Theorizing. *Academy of Management Discoveries*, *4*(4). https://doi.org/10.3929/ethz-b-000320207

Lamm, H., & Trommsdorff, G. (1973). Group versus individual performance on tasks requiring ideational proficiency (brainstorming): A review. *European Journal of Social Psychology*, *3*(4), 361–388. https://doi.org/10.1002/ejsp.2420030402

Leggett Dugosh, K., & Paulus, P. B [Paul B.] (2005). Cognitive and social comparison processes in brainstorming. *Journal of Experimental Social Psychology*, *41*(3), 313–320. https://doi.org/10.1016/j.jesp.2004.05.009

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*.

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, *120*, 262–273. https://doi.org/10.1016/j.jbusres.2020.07.045

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

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Nijstad, B. A., Dreu, C. K. W. de, Rietzschel, E. F., & Baas, M. (2010). The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European Review of Social Psychology*, *21*(1), 34–77. https://doi.org/10.1080/10463281003765323

OpenAI [Computer software]. www.openai.com

OpenAI Documentation. https://beta.openai.com/docs/api-reference/completions/create#completions/create-temperature

Osborn, A. F. (1953). *Applied Imagination: Principles and Procedures of Creative Thinking*. Scribner.

Paulus, P. B [Paul B.], Kohn, N. W., Arditti, L. E., & Korde, R. M. (2013). Understanding the Group Size Effect in Electronic Brainstorming. *Small Group Research*, *44*(3), 332–352. https://doi.org/10.1177/1046496413479674

Pinsonneault, A., Barki, H., Gallupe, R. B., & Hoppen, N. (1999). Electronic Brainstorming: The Illusion of Productivity. *Information Systems Research*, *10*(2), 110–133. https://doi.org/10.1287/isre.10.2.110

Ritter, S. M., & Mostert, N. M. (2018). How to facilitate a brainstorming session: The effect of idea generation techniques and of group brainstorm after individual brainstorm. *Creative Industries Journal*, *11*(3), 263–277. https://doi.org/10.1080/17510694.2018.1523662

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.

Siemon, D., Eckardt, L., & Robra-Bissantz, S. (2015). Tracking Down the Negative Group Creativity Effects with the Help of an Artificial Intelligence-Like Support System. In *2015 48th Hawaii International Conference on System Sciences* (pp. 236–243). IEEE. https://doi.org/10.1109/HICSS.2015.37

Sio, U. N., Kotovsky, K., & Cagan, J. (2015). Fixation or inspiration? A meta-analytic review of the role of examples on design processes. *Design Studies*, *39*, 70–99. https://doi.org/10.1016/j.destud.2015.04.004

Specht, A. R., & Buck, E. B. (2019). Crowdsourcing Change: An Analysis of Twitter Discourse on Food Waste and Reduction Strategies. *Journal of Applied Communications*, *103*(2). https://doi.org/10.4148/1051-0834.2240

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*.

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. (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