

# Influence of Solution Efficiency and Valence of Instruction on Additive and Subtractive Solution Strategies in Humans and GPT-4

Lydia Uhler<sup>\*</sup> <sup>a,b</sup>, Verena Jordan<sup>\*</sup> <sup>a</sup>, Jürgen Buder <sup>c</sup>, Markus Huff <sup>a,c</sup>, and Frank Papenmeier <sup>a</sup>

<sup>a</sup> Department of Psychology, Universität of Tübingen, Tübingen 72076, Germany

<sup>b</sup> Department of Psychology, Universität of Münster, Münster 48149, Germany

<sup>c</sup> Leibniz-Institut für Wissensmedien, Tübingen 72076, Germany

\* Both authors contributed equally to this work

Email: luhler@uni-muenster.de

Note: This is a preprint of our work that has not yet been peer-reviewed

(Last modified: 24 April 2024)

## **Abstract**

We explored the addition bias, a cognitive tendency to prefer adding elements over removing them to alter an initial state or structure, by conducting four preregistered experiments examining the problem-solving behavior of both humans and OpenAI's GPT-4 large language model. The experiments involved 588 participants from the U.S. and 680 iterations of the GPT-4 model. The problem-solving task was either to create symmetry within a grid (Experiments 1 and 3) or to edit a summary (Experiments 2 and 4). As hypothesized, we found that overall, the addition bias was present. Solution efficiency (Experiments 1 and 2) and valence of the instruction (Experiments 3 and 4) played important roles. Human participants were less likely to use additive strategies when subtraction was relatively more efficient than when addition and subtraction were equally efficient. GPT-4 exhibited the opposite behavior, with a strong addition bias when subtraction was more efficient. In terms of instruction valence, GPT-4 was more likely to add words when asked to "improve" compared to "edit", whereas humans did not show this effect. When we looked at the addition bias under different conditions, we found more biased responses for GPT-4 compared to humans. Our findings highlight the importance of considering comparable and sometimes superior subtractive alternatives, as well as reevaluating one's own and particularly the language models' problem-solving behavior.

## **Significance Statement**

Large language models (LLMs) are playing an increasingly supportive role in human decision-making and problem-solving, which are essential components of everyday life. Humans are subject to an addition bias during these processes, meaning they tend to add elements rather than subtract them from ideas, objects, or situations. The present study found that the addition bias exists not only in humans but also in OpenAI's GPT-4 LLM. The model showed a tendency to favor additive solution strategies even when subtraction would have been more efficient, and also when asked to "improve" a text, rather than "edit" it. These findings suggest a need to reconsider the use of subtractive solutions in daily life and to carefully evaluate the outputs of LLMs.

## Introduction

The integration of large language models (LLMs), such as OpenAI's Chat Generative Pre-Trained Transformer (ChatGPT), had a significant impact on the collaboration between humans and intelligent systems. Applications range from software development and testing to writing poetry, essays, business letters, or contracts (1). The impact of LLMs is particularly evident in process optimization and decision-making contexts (2) due to the models' advanced capabilities to make inferences and provide practical advice to users (3). However, it is important to acknowledge that both humans and LLMs are prone to biases (4, 5).

Cognitive biases are defined as systematic and erroneous response patterns in judgment and decision-making (6). Adams et al. (7) conducted a series of experiments to investigate a novel cognitive bias called addition bias, in human participants. This was tested in situations in which problems could be solved by either adding or removing elements. Additive transformations imply that the changed state has more elements than the original state, whereas subtractive transformations imply that the changed state has fewer elements than the original state (8). A consistent finding from Adams et al. (7) was that when people change ideas, objects, or situations, they tend to add rather than remove elements when solving various problems, such as stabilizing a Lego construction, improving a miniature golf course, creating symmetry within a grid, or rewriting an article summary. Moreover, studies by Adams et al. (7) showed a) that human participants even preferred adding elements in situations where a subtractive solution would have required fewer steps than additive solutions, and b) that instructions to "improve" a design led to a stronger addition bias than instructions to "worsen" a design.

According to Adams et al. (7), the high rate of additive transformations occurs because individuals generate both additive and subtractive ideas, but preferentially choose additive ones. In other words, people are prone to apply a "what can we add here?" heuristic, a default strategy that people use to simplify and speed up decision-making (9). Personal and situational factors trigger the use of heuristics (10), such as cognition or socio-cultural influences. The frequent use and perceived success of a particular strategy can make it more cognitively accessible, e.g., prior use of base rates enhanced their subsequent use and salience (11). Additive ideas have the potential to enhance recognition by others, e.g., policy initiatives to "plant a billion trees" might sound more appealing than "allowing forests to regenerate" (12). Other problem-solving biases, such as the base rate fallacy (13), sunk cost fallacy (14), or loss aversion (15) may also provide an explanation.

In their study of French adults, Fillion et al. (16) performed a replication of three out of the eight experiments of Adams et al. (7), thereby confirming the findings of the addition bias. Winter et al. (17) demonstrated that the addition bias is not limited to behavioral manifestations, but also extends to language. Frequency analysis of the Corpus of Contemporary American English (18) indicated that words associated with increasing quantity, such as "add" or "more", were more prevalent in English than those associated with decreasing quantity, such as "subtract" or "less". Furthermore, Winter et al. (17) showed that the addition bias carried over to OpenAI's GPT-3 LLM, the third generation of the GPT series, as observed in cloze probability analyses. These findings highlight the importance of language in the observed bias. LLMs are pre-trained on large amounts of human data. The models generate text sequentially, assigning probabilities to words based on context (19). During this process, human bias may be reflected due to various contributing factors. These factors include the training data with its source material and selection process (20), the algorithms used for data processing and learning (21), and the subjective judgments of human annotators who provide labels for the training data while fine-tuning the model's behavior (22).

In the present study, through a series of four experiments, we directly assessed the prevalence of the addition bias under specific conditions in human participants and in GPT-4, the latest LLM of

OpenAI. We preregistered that the number of iterations for GPT-4 would match the human sample size. The temperature parameter was kept at 0.7, which resulted in variability in the LLM's output. We used two different tasks to improve the generalizability of our findings. The symmetry task required the agents, humans and GPT-4, to create symmetry within a 4x4 grid by adding or removing as few crosses as possible. The summary task involved modifying the summary of an article by adding or removing words with the constraint that the final text had to contain a number of words that was outside a "forbidden" word count range. We manipulated either the efficiency of the solution strategy (Experiments 1 and 2) or the valence of the task instructions (Experiments 3 and 4). The assignment of tasks and manipulations to the four experiment is shown in Figure 1.

We manipulated the efficiency of a solution strategy through the number of steps required to reach the goal. Our experiments were designed to either prompt both additive and subtractive transformations equally (e.g., adding or deleting two elements to achieve the goal state, respectively) or to prompt subtractive transformations more strongly (e.g., adding three elements vs. subtracting one element to achieve the goal state). In Adams et al. (7), subtraction cues led to less additive transformations than no cues in the control condition. Fillon et al. (16) replicated this finding. The research on addition bias in relation to the valence of instruction (positively vs. negatively connotated task goals) has shown mixed results. Studies demonstrated that perceptual processing is heightened for emotionally salient words compared to neutral words (23). While Adams et al. (7) found no significant differences in the shares of subtractive solution strategies between the task goals of improving vs. worsening the status quo, Fillon et al. (16) found that fewer subtractive solutions were generated in the improvement condition. In purely linguistic analyses, Winter et al. (17) showed that addition-related words are generally used in more positive contexts than subtraction-related words. Furthermore, verbs of improvement are even more often related to addition (vs. to subtraction) than neutral verbs of change.

Following the empirical findings, we hypothesized that 1) additive solution strategies are chosen significantly more often than subtractive solution strategies across all conditions, 2) the likelihood of choosing additive solution strategies is larger when the solution efficiency for addition and subtraction is equal than when the solution efficiency for subtraction is higher and 3) the likelihood of choosing additive solution strategies is larger when the valence of the instruction is positive than when the valence of the instruction is neutral. As an additional research question, we investigated for potential differences between humans and GPT-4 with regard to the use of additive and subtractive solution strategies.

Understanding the addition bias is crucial because it can limit the consideration of alternative solutions and lead to the overuse of additive strategies by both, humans and LLMs. The consequences can be numerous, ranging from information overload (24) to policy and regulatory overload (25), the development of slow and bloated software (26), or complex computer models with increasing numbers of parameters (27).

## Results

**Overall addition bias.** A chi-square test was conducted to analyze the association between the type of solution strategy chosen (additive vs. subtractive) and the observed frequencies across all four experiments. The result was statistically significant [ $\chi^2(1, N = 1,268) = 104.49, P < .001$ ], indicating an association between the chosen solution strategy and the observed frequencies. Additive solution strategies (64.4%; 816/1268) were observed more frequently than subtractive strategies (35.6%; 452/1268).

A Pearson's chi-squared test was performed to assess homogeneity, specifically to examine the relationship between the selected solution strategy and the agent (humans vs. GPT-4). The result was statistically significant [ $\chi^2(1, N = 1,268) = 26.04, P < .001$ —Table 1], indicating an association between chosen solution strategy and agent, with humans being relatively less likely to choose

additive strategies (57.0%; 335/588) compared to GPT-4 (70.7%, 481/680). The Phi coefficient calculation resulted in  $\Phi = .14$ , suggesting a weak association between the chosen solution strategy and the agent, as per Cohen (28).

**Experiments 1–4: Factors influencing the addition bias.** For Experiment 1 to 4, we used generalized linear models (implemented in R with ref. 24–26) with the chosen solution strategy as the outcome variable. The agent and either the solution efficiency (in Experiments 1 and 2) or the valence of the task instruction (in Experiments 3 and 4) were used as predictors, along with their interaction. Solutions were included in the analysis if and only if the original state of the grid or summary was modified by adding or subtracting crosses or words, regardless of whether symmetry was achieved or the summary was outside the forbidden word count range.

An exploratory analysis using chi-square tests was conducted to assess whether additive strategies were chosen relatively more frequently than subtractive strategies for each experimental condition.

**Experiment 1: Solution efficiency in symmetry task.** The human participants and GPT-4 attempted to achieve symmetry while the efficiency of the solution strategy was manipulated through varying the initial number of crosses: either 6 crosses in the condition “addition and subtraction equally efficient” or 5 crosses in the condition “subtraction more efficient”. In both conditions, the instruction’s valence was neutral, indicated by the word “change”.

The results showed a significant interaction between agent and solution efficiency regarding the chosen solution strategy [ $\chi^2(1, n = 302) = 20.54, P < .001$  – see Fig. 2], a significant main effect of agent [ $\chi^2(1, n = 302) = 6.02, P = .01$ ], and a non-significant main effect of solution efficiency [ $\chi^2(1, n = 302) = 1.67, P = .20$ ]. While more efficient subtraction led to less additive solutions for humans as predicted [65.7% vs. 29.2%;  $\chi^2(1, n = 132) = 19.95, P < .001$ ], we observed the opposite effect for GPT-4, namely more efficient subtraction leading to more additive solutions [54.1% vs. 69.4%;  $\chi^2(1, n = 170) = 4.27, P = .04$ ].

The exploratory analysis of the chosen solution strategy for each condition (see Table S1 in the SI Appendix) showed significantly more additive than subtractive solutions for humans when addition and subtraction were equally efficient, and GPT-4 when subtraction was more efficient. No significant preference for one solution strategy occurred for GPT-4, when addition and subtraction were equally efficient. Significantly more subtractive than additive solutions occurred for humans when subtraction was more efficient.

**Experiment 2: Solution efficiency in summary task.** The objective of the task was to edit the summary having access to both the original summary and the corresponding article. Solution efficiency was varied by specifying a different forbidden word count range for the final summary. The lower bound was either similarly far away from the original word count as the upper bound (addition and subtraction equally efficient) or the lower bound was closer to the original word count than the upper bound (subtraction more efficient). In both conditions, the instruction’s valence was neutral, indicated by the word “edit”.

As in Experiment 1, the logistic regression model revealed a significant interaction between the agent and solution efficiency regarding the chosen solution strategy [ $\chi^2(1, n = 341) = 10.84, P < .001$  – see Fig. 2], a significant main effect of agent [ $\chi^2(1, n = 341) = 50.30, P < .001$ ], and a non-significant main effect of solution efficiency [ $\chi^2(1, n = 341) = 0.32, P = .57$ ]. We again observed that while more efficient subtraction led to less additive solutions for humans as predicted [57.5% vs. 39.3%;  $\chi^2(1, n = 171) = 5.78, P = .02$ ], the opposite effect also again

occurred for GPT-4, namely more efficient subtraction leading to more additive solutions [77.6% vs. 90.6%;  $\chi^2(1, n = 170) = 5.44, P = .02$ ].

The exploratory analysis of the chosen solution strategy for each condition (see Table S1 in the SI Appendix) showed significantly more additive than subtractive solutions for GPT-4 in both conditions. For humans, no significant preference for one solution strategy occurred when addition and subtraction were equally efficient, and significantly more subtractive than additive solutions occurred when subtraction was more efficient.

**Experiment 3: Valence of instruction in symmetry task.** To manipulate the valence of the instruction in the symmetry task, the agent was instructed to either “change” or “improve” the given grid to achieve the goal of creating symmetry. In both conditions, the efficiency of the solution strategy was the same for both addition and subtraction, that is, 6 crosses in the original grid.

The logistic regression model revealed a non-significant interaction between the agent and the valence of instruction regarding the chosen solution strategy [ $\chi^2(1, n = 289) = 0.95, P = .33$  – see Fig. 2], a marginally significant main effect of agent [ $\chi^2(1, n = 289) = 3.85, P = .05$ ], and a non-significant effect of the valence of instruction [ $\chi^2(1, n = 289) = 0.62, P = .43$ ]. Thus, we observed no effect of the valence of instruction on the chosen solution strategy for both humans and GPT-4 in the spatial symmetry task.

The exploratory analysis of the chosen solution strategy for each condition (see Table S1 in the SI Appendix) showed significantly more additive than subtractive solutions for humans when the valence of instruction was neutral. For all other condition, there was no significant preference for one solution strategy. Please note that despite one condition of humans being just significant and the other being just not significant for this exploratory analysis, this difference in significance levels should not be interpreted because there was no significant effect of valence of instruction on chosen solution strategy in this experiment (see above).

**Experiment 4: Valence of instruction in summary task.** To manipulate the valence of instruction in the summary task, the agent was directed to either “edit” or “improve” the provided article summary. In both conditions, the efficiency of the solution strategy was the same for both addition and subtraction, that is, the lower bound of the forbidden word count range was similarly far away from the original word count as the upper bound.

The logistic regression model showed a significant interaction between the agent and the valence of instruction regarding the chosen solution strategy [ $\chi^2(1, n = 336) = 6.84, P < .01$  – see Fig. 2], and significant main effects for both the agent [ $\chi^2(1, n = 336) = 15.20, P < .001$ ], and the valence of instruction [ $\chi^2(1, n = 336) = 10.59, P < .01$ ]. While valence of instruction did not affect the chosen solution strategy for humans [65.1% vs. 73.5%;  $\chi^2(1, n = 166) = 1.38, P = .24$ ], we observed the predicted increase in additive solutions for positive valence of instruction for GPT-4 [76.5% vs. 96.5%;  $\chi^2(1, n = 170) = 15.70, P < .001$ ].

The exploratory analysis of the chosen solution strategy for each condition (see Table S1 in the SI Appendix) showed significantly more additive than subtractive solutions for all conditions.

**Exploratory: Proportion goal achieved.** The target of our study was to analyze the solutions generated by humans and the GPT-4 LLM to investigate the presence of addition bias, regardless of whether the goal of creating symmetry or a summary outside the forbidden word count range was achieved. As a further exploratory analysis, we examined the rates to which these task goals were achieved (see Table S2 in the SI Appendix for details). For the symmetry task, humans

consistently obtained high rates of full symmetry across all experimental conditions (80.0%–97.0%), while GPT-4 consistently achieved low rates of full symmetry (1.2%–3.5%). For the summary task, humans obtained high rates of achieving the goal by exceeding or falling below the forbidden word count range in their final summary in all experimental conditions (81.9%–86.9%), while GPT-4 had low rates of exceeding or falling below the forbidden word count range (15.3%–17.6%), except for one condition (67.1%). This exception was the “improve” condition of the summary task, which was a condition where GPT-4 exhibited a very strong addition bias likely leading to the word count exceeding the forbidden word range without GPT-4 actually understanding the task.

## Discussion

The concept of addition bias is reflected in the phrase “the more, the better”, which implies a preference for additive solution strategies over subtractive solution strategies (7). Is it recognized that sometimes “less is more”, especially when subtraction is more efficient than addition? Does the language used in instructions influence the choice of solution strategies? As stated in the introduction, language models like ChatGPT are becoming more commonly used for problem-solving (2). What strategies is the GPT-4 LLM using? Does the model exhibit similar solution behavior to that of humans? How does it generally perform in the tasks? We aimed to answer these questions through this experimental series.

The first research objective of this study was to further investigate human addition bias and to replicate previous research. Consistent with our preregistered first hypothesis, we found an overall bias towards addition. That is, additive solution strategies were chosen significantly more often than subtractive solution strategies. Interestingly, we also observed that this overall addition bias was stronger for GPT-4 than for humans.

The manipulation of solution efficiency influenced the likelihood of choosing additive solution strategies for both spatial and verbal problems (Experiments 1 and 2). However, this influence of solution efficiency on addition bias worked in opposite directions for humans and GPT-4. Supporting the second hypothesis, we found that humans were more likely to choose additive solution strategies when addition and subtraction were equally efficient compared to when subtraction was more efficient. However, the results for GPT-4 were contrary to the hypothesis, showing a higher likelihood of choosing additive solution strategies when subtraction was more efficient than when addition and subtraction were equally efficient.

The manipulation of the valence of instruction had no significant effect on the addition bias in the spatial problem for both agents (Experiment 3), contrary to our third hypothesis. In the linguistic problem (Experiment 4), however, a positive valence of instruction (“improve”) as compared to a neutral valence of instruction (“edit”) caused a significant increase in addition bias for GPT-4 with no effect for humans. Thus, our third hypothesis was only supported for GPT-4 in the linguistic problem.

Our study showed a tendency towards addition rather than subtraction, which is in line with previous research (7, 16, 17). The findings partially supported research indicating that LLMs and humans exhibit similar patterns in heuristic use and bias occurrence across various tasks (32). This supports the notion that human biases are incorporated into LLMs through data or training, and that the models can be considered “stochastic parrots” that reproduce biases in their outputs (33). GPT-4 exhibited a higher sensitivity to certain words, such as “improve”, and failed to recognize the superior efficiency of the subtractive solution. As a result, it heavily relied on additive solution strategies for both scenarios. Although this LLM may accurately reproduce language, it failed to provide a deep and authentic understanding. This is also reflected in the high rates of unfulfilled full symmetry and inaccurate summary lengths. The LLM outputs provide evidence of the consequences of processes that produce additive bias. However, the LLM does

not show the cognitive and emotional processes that are thought to be critical to this bias (34). Therefore, it is worth pursuing further investigation into the role of language in producing addition bias in humans.

The participants showed an addition bias when all solutions were equally efficient. This means that only for humans, the likelihood of providing subtractive responses increased when subtraction was perceived as the less time-consuming and effort-intensive option. Thus, it appears that humans often use the “what can we add here?” heuristic as a standard tactic for simplifying and speeding up decision making (9). But in a scenario of efficient subtractive solutions, this heuristic can be overcome as people expend additional cognitive effort to think of other, more efficient answers. We found it surprising that in GPT-4, the relatively higher efficiency of subtraction led to an increased use of additive strategies, indicating an increased addition bias. One possible explanation for the result is the different baseline number of crosses (6 in the “addition and subtraction equally efficient” condition vs. 5 in the “subtraction more efficient” condition). Research showed that GPT-3 can be influenced by random anchors during estimation (32). It is possible that GPT-4 perceived a higher potential for addition when there were 5 crosses in 16 boxes, compared to 6 crosses in 16 boxes. It was found – at least in some instances – that AI-based decision-making can magnify biases and create new classifications and criteria, leading to new types of biases (35). This is critical, because human users may unknowingly adopt these automated biases, known as the automation bias (36), which had been observed in clinical and HR processes (37, 38).

In line with the results of Adams et al. (7) but in contrast to those of Winter et al. (17), there was no effect of the valence of instruction on human solution strategy choices in either the spatial or the linguistic task. This may imply that despite the existence of differences between verbs of change and verbs of improvement in their relation to addition in the English language (17), such a subtle task variation does not suffice to trigger different behavioral outcomes in humans.

In addition to the comparison within the conditions of solution efficiency and within the conditions of valence of instruction, we also examined the predictions put forth by Adams et al. (7). This included a relatively higher frequency of additive solution strategies compared to subtractive strategies, despite the potential for subtraction to be more efficient. In contrast to Adams et al. (7), we did not find the addition bias in every condition. When it was more efficient to choose subtractive solution strategies, for example, no addition bias could be found in humans. However, it should be noted that we performed a linguistic adaption of the symmetry task, which might have provided an additional cue for subtraction as compared to the original grid task. More precisely, we exchanged the verb “toggle” by the phrase “switch any field from ‘[X]’ to ‘[ ]’ or from ‘[ ]’ to ‘[X]’”. Despite this potential subtraction cue, however, addition bias was still found in humans when the solution efficiency was equal for additive and subtractive solutions. As our study did not include any no-cue control group, the effect of this potential subtraction cue – especially in conjunction with the variation of solution efficiency – cannot be further explored. Similarly, operationalizing the solution efficiency manipulation in the summary task by specifying prohibited word count ranges also provided an originally non-existent cue for subtractive solutions. Given the significant impact of subtraction cues found by Adams et al. (7), these task deviations might explain the differences in the results.

Our findings support the notion that human decision-making is influenced by cognitive or cultural processes, as demonstrated in previous research (7). It is debatable whether these processes result in the addition bias, or if other problem-solving biases, such as the base rate fallacy (13), sunk cost fallacy (14), or loss aversion (15), provide a more accurate explanation. These biases overlook general probabilities, the benefits of an investment stopping, or the similarity of losses to equivalent gains. Future studies should aim to directly control for the different biases or measure their relative importance and how they may combine. With respect to the stated biases, it is important to consider the counterpart of the addition bias, called the subtraction neglect. The

preference for additive solutions may be reinforced by the fact that subtractive solutions are less likely to be recognized or valued (9).

Through this experimental series, we gained a better understanding of the addition bias and its prerequisites from a comparative perspective. However, further research is needed to address several points that are still up for discussion.

First of all, the addition bias was shown to occur in humans and LLMs, especially under certain conditions. Further investigation into the mechanisms of human thinking and the functioning of LLM algorithms is needed. To gain a better understanding of why the addition bias occurs, future research can use specific methods from cognitive psychology, such as think-aloud protocols.

Research showed that thinking aloud can aid in understanding human cognition and behavior (39). This can help determine whether perception, recognition, or other heuristic processes are critical factors leading to additive changes. ChatGPT demonstrated the ability to provide reliable and high-quality explanations for its predictions (40), which is required for further research.

Prompt engineering and in particular chain-of-thought prompting (COT) offer a promising way to improve the understanding of LLMs such as ChatGPT. Querying intermediate steps during the problem-solving process can provide insight into the decision-making processes of the model (41).

Secondly, the sample size of participants did not meet the preregistered size. We assume that this was mainly due to participants' difficulties in understanding the instructions, as evidenced by this response: "I would argue that the solution does not exist due to the fact X can only move to an empty box. It cannot be duplicated. [...]" – case 232, Experiment 1. According to the principles of Gestalt psychology, it could be inferred that difficulties arose in distinguishing between the ground and figure, as Juvrud et al. (42) suggested with regard to the grid task designed by Adams et al. (7). Providing examples can prevent misunderstandings in task functioning. For humans, general instructions and goal descriptions can aid in learning and transfer by requiring the learner to make an effort to comprehend the subject and utilize learning-friendly cognitive strategies (43). In further research on LLMs, few-shot learning could be advantageous compared to zero-shot learning. Few-shot learning involves training models on a small amount of labeled data, such as exemplary tasks and solutions, rather than relying solely on knowledge learned from other tasks, that is, zero-shot learning (44). Relying on few-shot learning, the potential for certain solution strategies to be influenced by the provision of solution examples represents a challenge that must be addressed.

GPT-4 had limited success in achieving the goals of full symmetry in the grid and generating a final summary outside the forbidden word count range. Due to the small number of cases, the preregistered exploratory analyses that only considered correct solutions were not conducted. Although LLMs have demonstrated impressive multi-modal capabilities (45, 46), spatial reasoning, which is required for symmetry generation, remains a significant challenge for the language models (47). Prompting strategies, such as Chain of Thought and Tree of Thought, which generalizes the Chain of Thought approach, can be used to improve spatial reasoning skills (47). Pre-trained LLMs exhibit a clear insensitivity to the contextual effects of negations (48). This may explain GPT-4's difficulty with certain instructions, such as "not between 68 and 94 words," as the model may not adequately account for the negated range and generate a summary within that prohibited range. Although fine-tuning the models on negative sentences can improve their performance, it is important to note that handling negation remains a persistent challenge for LLMs in terms of generalization and understanding (49). When experimenting with LLMs of this type, future studies could convert negative statements to positive affirmations. For example, instead of stating "not between 68 and 94 words", write "either less than 68 words or more than 94 words".

This study was conducted with OpenAI's GPT-4 LLM. Further research is required to determine if there is any additional bias in other LLMs and to identify the factors that contribute to this bias, including data, algorithms, and labeling by human annotators. Future investigations could also explore other dimensions of AI-based systems, such as creative tasks involving image generation. Adams et al. (7) demonstrated an addition bias in design tasks when participants were asked to add or remove elements to a mini-golf course. Our findings, which were based on a sample of participants from the U.S. with a good command of English, raise questions about linguistic generalizability. Previous research has highlighted the significance of language in comparative studies of American and Swedish adults (42). Further investigation is needed to fully understand the role of language as a potential moderator.

In the end, how can biases in humans and LLMs be addressed? There is much debate surrounding this question (50, 51). Heuristics have proven useful in some cases (52). Algorithmic bias is nothing less than the reproduction of biases that are inherently human (20), although sometimes in an amplified way, as seen in this study. If the decision is made to mitigate the effects of the addition bias replicated in this study, several strategies can be employed. One way to mitigate bias in decision-making is to increase awareness of its potential impact. By recognizing the tendency to prioritize additive solutions, users can consciously consider subtractive options and strive for a more balanced approach to problem solving. When working with an LLM, it is important to explicitly request the consideration of subtractive solutions and to critically reflect on the LLM's solutions as well as their accuracy. Otherwise, there is a risk of habitual acceptance of recommendations by LLMs (53). While defaulting to a search for additive ideas may often benefit users, it is not a reason to settle for acceptable additive changes provided by LLMs, other people, or oneself. Therefore, this is a call to consider comparable and sometimes superior subtractive alternatives, as well as to reevaluate one's own and LLM's solution behavior.

## Materials and Methods

**Participants and GPT-4.** The study was approved by the ethics committee of the Leibniz-Institut für Wissensmedien in Tübingen.

The human sample was collected through the online research platform Prolific ([www.prolific.com](http://www.prolific.com)) [initial survey: August 9-11, 2023, additional survey: September 7-10, 2023]. Inclusion criteria required subjects to have a good command of English and to be over 18 years of age. This setting was set to Prolific to avoid having to exclude any participants based on these criteria post hoc. A preliminary data filtering was carried out so that double participations were removed as preregistered. Furthermore, participants who did not consent to the study prerequisites or to the final submission of their data were not considered. During our pretest of the symmetry task, we observed responses that led us to believe that the participants were using external tools, such as LLMs, to complete the task. Therefore, we updated our preregistration before final data collection such that participants who lost focus on the experimental webpage during the task completion phase were excluded from the analysis and replaced. Loss of focus is defined as any scenario in which the experimental tab or window is not the active focus of the participant, e.g., when participants open other applications such as ChatGPT or the email inbox while working on the task. As technology continues to advance, the possibility of confounding variables arising from external tools that participants may use to complete experimental tasks has become increasingly important (54). Table S3 in the SI Appendix presents statistics on the number of participants excluded for each reason in each experiment. Additionally, we excluded participants who did not change the number of crosses or words from the original to the final grid or the final summary. Statistics on the number of excluded participants due to no change being made are also presented in Table S3 of the SI Appendix. This is because our main goal is to analyze addition bias, which requires additive or subtractive transformations to the original state.

Table S4 in the SI Appendix presents the demographic information for the final group of participants in the four experiments, including sample size, gender distribution, and age statistics. In our study, a total of 588 participants across four experiments were analyzed, predominantly identifying as male or female, with a mean age ranging from 34 to 37 years.

On August 10, 2023, the study was conducted using GPT-4, performing 680 iterations. GPT-4 is a large multimodal language model developed by OpenAI and the fourth in the GPT series (46). It was released on March 14, 2023, and is based on both publicly available data and data licensed from third-party providers. However, no further details regarding the architecture (including model size), hardware, dataset construction, training computation, or training method are provided. The transformer-based model GPT-4 was pre-trained to predict the next token and fine-tuned by developers using reinforcement learning from human feedback. Studies show that GPT-4 is an improvement over the previous version, GPT-3.5, e.g., in text summarization, dialog generation and in understanding of complex information (55, 56). Due to the model's high performance we chose GPT-4 for this study. The experiments were conducted using the following parameter values: The temperature parameter was maintained at 0.7, its default value. The maximum length of available tokens was set to 500, which was considered sufficient for GPT-4 to generate appropriate output for the given tasks. For both the symmetry task and the summary task, we used the "Zero-Shot" method (44). This method only involves providing instructions without examples. As with participants, we excluded GPT-4 iterations in which the number of crosses or words did not change from the initial to the final grid or summary. See Table S5 in the SI Appendix for statistics on the number of exclusions for this reason.

A power analysis was conducted using G\*Power 3.1 (57) to determine the sample sizes for the human samples. The effect size was specified in terms of the odds ratio. The odds ratios used for the calculation were based on the findings of Adams et al. (7). The power analysis indicated that a sample size of 170 is needed for the "solution efficiency" condition and a sample size of 166 is needed for the "valence of Instruction" condition to achieve 80% power, given the odds ratio of 2.47 in the "subtraction neglect" condition and the odds ratio of 0.41 in the "valence of instruction" condition. Further information on the calculation of the sample size is documented in the preregistration (see <https://osf.io/6pkwb>).

For each of the four experiments, we targeted a sample size of 170 participants and 170 iterations with GPT-4. We preregistered that in the event of participant exclusions, replacements were made until at least 170 participants were surveyed. We specified that the number of iterations for GPT-4 had to match the human sample size. Because of unexpected participant exclusions during the first survey, we conducted an additional survey based on the exclusion rates of the previous survey. Even after the additional survey, the number of participants in all experiments but Experiment 2 was still less than 170. We did not collect any additional data after the additional survey. The exclusions of the GPT-4 iterations in which no change to the original state occurred were replaced by additional iterations that we had already collected in the first survey of GPT-4 iterations.

**Design.** The study was conducted using an experimental design consisting of four experiments. Experiments 1 and 3 focused on solution efficiency in spatial and linguistic problems, while Experiments 2 and 4 focused on valence in spatial and linguistic problems. Additionally, the study included an agent factor to compare humans and GPT-4. In Experiments 1 and 2, we used a 2 (agent: human vs. GPT-4) x 2 (solution efficiency: efficiency of addition and subtraction equal vs. subtraction more efficient than addition) between-subject design. The instructional valence was neutral. In Experiments 3 and 4 we used a 2 (agent: human vs. GPT-4) x 2 (valence of instruction: neutral vs. positive) between-subject design. The efficiency of addition and subtraction was equal. Participants were randomly assigned to only one condition in each experiment to prevent learning or training effects (58). Regarding the iterations of GPT-4, the

prompts were given independently with a new prompt each time to prevent any training effect in terms of fine-tuning.

**Material.** For the data collection with the human participants, we used questionnaires that we created with SoSci-Survey. On the Prolific recruitment platform, we set the requirement that participants can only participate with a desktop computer, but not with a tablet or mobile phone. The questionnaires included two tasks: either a grid-based symmetry task or a summary task. Both tasks were adopted from Adams et al. (7). The grid within the symmetry task had to be adapted for implementation on SoSci-Survey and for input to GPT-4 in that the colored boxes were represented by “[X]” and the non-colored boxes were represented by empty brackets “[ ]”. For the summary task, the article by Adams et al. (7) was not suitable because their summary was created by study participants or GPT-3. Instead, an article comparable in terms of word count and its summary were taken from an English textbook. It is an article from the Australian Broadcasting Corporation entitled “Thor’ aircraft, capable of dropping 15,000 liters of water, arrives in NSW ahead of bushfire season” written by Nick Dole (59) and published in 2015. Vignettes of the symmetry task and the summary task, both in the “addition and subtraction equally efficient” condition with a neutral instruction valence, are provided in the SI Appendix.

**Procedure.** For the human data collection, participants were required to read and agree to the participation information and data protection statement before each experiment. If the participant did not agree to this, the study ended without entitlement to remuneration. If the person agreed, the demographic data was entered. The task then appeared. Once the task had been completed, the study participants were given the opportunity to withdraw their data. Regardless of the withdrawal, the participants then received the compensation code and were redirected back to Prolific. The compensation was £0.50 for the symmetry task and £1.00 for the summary task. The median completion times in the first and the additional survey of all four experiments are shown in Table S6. For the data collection with the GPT-4 LLM, the prompts for the symmetry and summary tasks were input into GPT-4 separately via an API.

**Analysis.** The study measured the categorical dependent variable, which is the type of solution strategy chosen, as either additive or subtractive. In the symmetry task, the final number of crosses within the grid’s boxes was used to determine whether the strategy was additive (more crosses than original grid) or subtractive (less crosses than original grid). In the summary task, the final number of words was used to determine whether the strategy was additive (more words than original summary) or subtractive (less words than original summary).

As described, we conducted chi-squared tests and logistic regression analyses. To predict the chosen solution strategy, we used a binomial generalized linear model with a logit link function.

For the coding of participants’ and GPT-4’s responses in the symmetry task, we used the same coding schema. We coded the final count of crosses within the grid’s boxes (called box count), the original box count, whether the final box count was unequal to the original box count, the solution strategy, the correctness of the solution, and whether it was the most efficient solution strategy within the chosen solution strategies, which is the strategy with the fewest steps. To code the participants’ and GPT-4’s responses in the summary task, we first calculated the final word count. Then, we applied the coding schema and coded whether the final word count was unequal to the original word count, the solution strategy, and whether the final word count was outside or within the forbidden word count range.

The data and analysis scripts are available at <https://osf.io/c78rm/>

## **Acknowledgments**

We thank Gerrit Anders for supporting us with collecting the GPT-4 iterations.

## References

1. E. Kasneci, *et al.*, ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences* **103**, 102274 (2023).
2. A. K. Kar, P. S. Varsha, S. Rajan, Unravelling the Impact of Generative Artificial Intelligence (GAI) in Industrial Applications: A Review of Scientific and Grey Literature. *Glob J Flex Syst Manag* **24**, 659–689 (2023).
3. H. Nori, N. King, S. M. McKinney, D. Carignan, E. Horvitz, Capabilities of GPT-4 on Medical Challenge Problems. arXiv [Preprint]. (2023). <https://doi.org/10.48550/ARXIV.2303.13375>.
4. A. Acerbi, J. M. Stubbersfield, Large language models show human-like content biases in transmission chain experiments. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2313790120 (2023).
5. M. Binz, E. Schulz, Using cognitive psychology to understand GPT-3. *Proc. Natl. Acad. Sci. U.S.A.* **120**, e2218523120 (2023).
6. A. Tversky, D. Kahneman, Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *Science* **185**, 1124–1131 (1974).
7. G. S. Adams, B. A. Converse, A. H. Hales, L. E. Klotz, People systematically overlook subtractive changes. *Nature* **592**, 258–261 (2021).
8. G. Vergnaud, “A Classification of Cognitive Tasks and Operations of Thought Involved in Addition and Subtraction Problems” in *Addition and Subtraction*, 1st Ed., T. P. Carpenter, J. M. Moser, T. A. Romberg, Eds. (Routledge, 2020), pp. 39–59.
9. T. Meyvis, H. Yoon, Adding is favoured over subtracting in problem solving. *Nature* **592**, 189–190 (2021).
10. C. Del Campo, S. Pauser, E. Steiner, R. Vetschera, Decision making styles and the use of heuristics in decision making. *J Bus Econ* **86**, 389–412 (2016).
11. Z. Ginossar, Y. Trope, Problem solving in judgment under uncertainty. *Journal of Personality and Social Psychology* **52**, 464–474 (1987).
12. N. Sallou, A. Stritih, In complexity we trust: learning from the socialist calculation debate for ecosystem management. *Environ. Res. Lett.* **18**, 051001 (2023).
13. D. Kahneman, A. Tversky, On the psychology of prediction. *Psychological Review* **80**, 237–251 (1973).
14. H. R. Arkes, P. Ayton, The sunk cost and Concorde effects: Are humans less rational than lower animals? *Psychological Bulletin* **125**, 591–600 (1999).
15. D. Kahneman, A. Tversky, Prospect Theory: An Analysis of Decision under Risk. *Econometrica* **47**, 263 (1979).
16. A. A. Fillon, F. Girandola, N. Bonnardel, L. Souchet, “People systematically overlook subtractive changes (2021): Replication and extension” (PsyArXiv, 2022).
17. B. Winter, M. H. Fischer, C. Scheepers, A. Myachykov, More is Better: English Language Statistics are Biased Toward Addition. *Cognitive Science* **47**, e13254 (2023).

18. M. Davies, The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* **25**, 447–464 (2010).
19. A. Vaswani, *et al.*, Attention is All you Need in *Advances in Neural Information Processing Systems*, I. Guyon, *et al.*, Eds. (Curran Associates, Inc., 2017).
20. A. Caliskan, J. J. Bryson, A. Narayanan, Semantics derived automatically from language corpora contain human-like biases. *Science* **356**, 183–186 (2017).
21. S. L. Blodgett, S. Barocas, H. Daumé III, H. Wallach, Language (Technology) is Power: A Critical Survey of “Bias” in NLP in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (Association for Computational Linguistics, 2020), pp. 5454–5476.
22. D. Hovy, S. Prabhumoye, Five sources of bias in natural language processing. *Language and Linguist. Compass* **15**, e12432 (2021).
23. F. M. M. Citron, M. A. Gray, H. D. Critchley, B. S. Weekes, E. C. Ferstl, Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia* **56**, 79–89 (2014).
24. P. G. Roetzel, Information overload in the information age: a review of the literature from business administration, business psychology, and related disciplines with a bibliometric approach and framework development. *Bus Res* **12**, 479–522 (2019).
25. M. Hinterleitner, C. Knill, Y. Steinebach, The growth of policies, rules, and regulations: A review of the literature and research agenda. *Regulation & Governance* rego.12511 (2023). <https://doi.org/10.1111/rego.12511>.
26. J. D. Bayliss, The Data-Oriented Design Process for Game Development. *Computer* **55**, 31–38 (2022).
27. U. Proske, S. Ferrachat, D. Neubauer, M. Staab, U. Lohmann, Assessing the potential for simplification in global climate model cloud microphysics. *Atmos. Chem. Phys.* **22**, 4737–4762 (2022).
28. J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, 2nd Ed. (Routledge, 1988).
29. J. Fox, S. Weisberg, *An R companion to applied regression*, Third edition (SAGE, 2019).
30. H. Wickham, R. François, L. Henry, K. Müller, dplyr: A Grammar of Data Manipulation. (2023). Deposited 2023.
31. H. Wickham, J. Bryan, readxl: Read Excel Files. (2022). Deposited 2022.
32. G. Suri, L. R. Slater, A. Ziaeef, M. Nguyen, Do Large Language Models Show Decision Heuristics Similar to Humans? A Case Study Using GPT-3.5. arXiv [Preprint]. (2023). <https://doi.org/10.48550/ARXIV.2305.04400>.
33. E. M. Bender, T. Gebru, A. McMillan-Major, S. Shmitchell, On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?  in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (ACM, 2021), pp. 610–623.
34. S. Trott, C. Jones, T. Chang, J. Michaelov, B. Bergen, Do Large Language Models Know What Humans Know? *Cognitive Science* **47**, e13309 (2023).

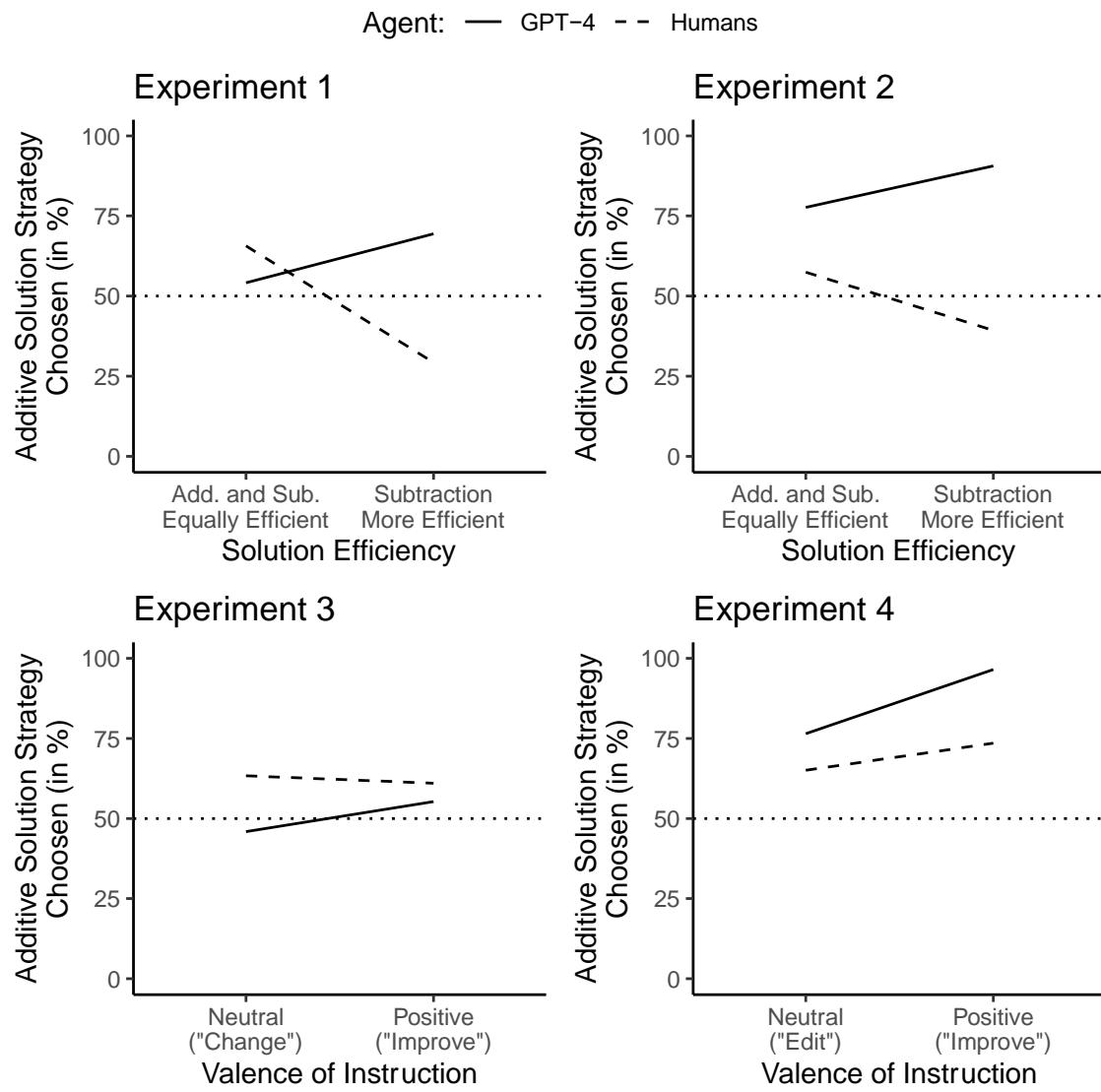
35. E. Ntoutsi, *et al.*, Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining Knowl Discov* **10** (2020).
36. L. J. Skitka, K. L. Mosier, M. Burdick, Does automation bias decision-making? *International Journal of Human-Computer Studies* **51**, 991–1006 (1999).
37. K. Goddard, A. Roudsari, J. C. Wyatt, Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J Am Med Inform Assoc* **19**, 121–127 (2012).
38. A. Köchling, M. C. Wehner, Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Bus Res* **13**, 795–848 (2020).
39. K. A. Ericsson, “Protocol Analysis” in *A Companion to Cognitive Science*, 1st Ed., W. Bechtel, G. Graham, Eds. (Wiley, 2017), pp. 425–432.
40. B. Li, *et al.*, Evaluating ChatGPT’s Information Extraction Capabilities: An Assessment of Performance, Explainability, Calibration, and Faithfulness. arXiv [Preprint]. (2023). <https://doi.org/10.48550/ARXIV.2304.11633>.
41. J. Wei, *et al.*, Chain-of-Thought Prompting Elicits Reasoning in Large Language Models in *Advances in Neural Information Processing Systems*, S. Koyejo, *et al.*, Eds. (Curran Associates, Inc., 2022), pp. 24824–24837.
42. J. Juvrud, L. Myers, P. Nyström, People overlook subtractive changes differently depending on age, culture, and task. *Sci Rep* **14**, 1086 (2024).
43. E. Eiriksdottir, R. Catrambone, Procedural Instructions, Principles, and Examples: How to Structure Instructions for Procedural Tasks to Enhance Performance, Learning, and Transfer. *Hum Factors* **53**, 749–770 (2011).
44. T. B. Brown, *et al.*, Language Models Are Few-Shot Learners in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS’20., (Curran Associates Inc., 2020).
45. D. Zhu, J. Chen, X. Shen, X. Li, M. Elhoseiny, MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. (2023). <https://doi.org/10.48550/ARXIV.2304.10592>.
46. OpenAI, *et al.*, GPT-4 Technical Report. (2023). <https://doi.org/10.48550/ARXIV.2303.08774>.
47. F. Li, D. C. Hogg, A. G. Cohn, Advancing Spatial Reasoning in Large Language Models: An In-Depth Evaluation and Enhancement Using the StepGame Benchmark. arXiv [Preprint]. (2024). <https://doi.org/10.48550/ARXIV.2401.03991>.
48. A. Ettinger, What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models. *Transactions of the Association for Computational Linguistics* **8**, 34–48 (2020).
49. I. García-Ferrero, B. Altuna, J. Alvez, I. Gonzalez-Dios, G. Rigau, This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, (Association for Computational Linguistics, 2023), pp. 8596–8615.
50. G. Gigerenzer, W. Gaissmaier, Heuristic Decision Making. *Annu. Rev. Psychol.* **62**, 451–

- 482 (2011).
51. E. Ferrara, Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models. arXiv [Preprint]. (2023). <https://doi.org/10.48550/ARXIV.2304.03738>.
  52. S. Hart, Adaptive Heuristics. *Econometrica* **73**, 1401–1430 (2005).
  53. E. A. M. Van Dis, J. Bollen, W. Zuidema, R. Van Rooij, C. L. Bockting, ChatGPT: five priorities for research. *Nature* **614**, 224–226 (2023).
  54. T. Heyman, G. Heyman, The impact of ChatGPT on human data collection: A case study involving typicality norming data. *Behav Res* (2023). <https://doi.org/10.3758/s13428-023-02235-w>.
  55. N. Oh, G.-S. Choi, W. Y. Lee, ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Ann Surg Treat Res* **104**, 269 (2023).
  56. Y. Liu, *et al.*, Summary of ChatGPT-Related research and perspective towards the future of large language models. *Meta-Radiology* **1**, 100017 (2023).
  57. F. Faul, E. Erdfelder, A. Buchner, A.-G. Lang, Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods* **41**, 1149–1160 (2009).
  58. G. Charness, U. Gneezy, M. A. Kuhn, Experimental methods: Between-subject and within-subject design. *Journal of Economic Behavior & Organization* **81**, 1–8 (2012).
  59. N. Dole, “Thor” aircraft, capable of dropping 15,000 litres of water, arrives in NSW ahead of bushfire season. *ABC News* (2015). Available at: <https://www.abc.net.au/news/2015-09-01/thor-water-bombing-aircraft-arrives-in-nsw-ahead-bushfire-season/6740796> [Accessed 1 July 2023].

## Figures and Tables

Symmetry Task		Summary Task																																									
<p>"There is a digital 4-by-4 grid [...]. You can switch any field from [X] to [ ] or from [ ] to [X]."</p>		<p>There is an article and a summary of it. The summary has 78 words.</p>																																									
<b>Experiment 1</b>		<b>Experiment 2</b>																																									
<b>Addition and subtraction equally efficient</b>	<b>Subtraction more efficient</b>	<b>Addition and subtraction equally efficient</b>	<b>Subtraction more efficient</b>																																								
<table border="0"> <tr> <td>A</td><td>B</td><td>C</td><td>D</td></tr> <tr> <td>1 [X]</td><td>[ ]</td><td>[ ]</td><td>[ ]</td></tr> <tr> <td>2 [ ]</td><td>[X]</td><td>[X]</td><td>[ ]</td></tr> <tr> <td>3 [ ]</td><td>[X]</td><td>[X]</td><td>[ ]</td></tr> <tr> <td>4 [X]</td><td>[ ]</td><td>[ ]</td><td>[ ]</td></tr> </table>	A	B	C	D	1 [X]	[ ]	[ ]	[ ]	2 [ ]	[X]	[X]	[ ]	3 [ ]	[X]	[X]	[ ]	4 [X]	[ ]	[ ]	[ ]	<table border="0"> <tr> <td>A</td><td>B</td><td>C</td><td>D</td></tr> <tr> <td>1 [X]</td><td>[ ]</td><td>[ ]</td><td>[ ]</td></tr> <tr> <td>2 [ ]</td><td>[X]</td><td>[X]</td><td>[ ]</td></tr> <tr> <td>3 [ ]</td><td>[X]</td><td>[X]</td><td>[ ]</td></tr> <tr> <td>4 [ ]</td><td>[ ]</td><td>[ ]</td><td>[ ]</td></tr> </table>	A	B	C	D	1 [X]	[ ]	[ ]	[ ]	2 [ ]	[X]	[X]	[ ]	3 [ ]	[X]	[X]	[ ]	4 [ ]	[ ]	[ ]	[ ]	<p>"edit this summary"</p> <p>"The edited summary MUST NOT be between 63 and 93 words long."</p>	<p>"edit this summary"</p> <p>"The edited summary MUST NOT be between 68 and 98 words long."</p>
A	B	C	D																																								
1 [X]	[ ]	[ ]	[ ]																																								
2 [ ]	[X]	[X]	[ ]																																								
3 [ ]	[X]	[X]	[ ]																																								
4 [X]	[ ]	[ ]	[ ]																																								
A	B	C	D																																								
1 [X]	[ ]	[ ]	[ ]																																								
2 [ ]	[X]	[X]	[ ]																																								
3 [ ]	[X]	[X]	[ ]																																								
4 [ ]	[ ]	[ ]	[ ]																																								
<p>"change [...] so that the grid is perfectly symmetrical"</p>		<p>"change [...] so that the grid is perfectly symmetrical"</p>																																									
<b>Experiment 3</b>		<b>Experiment 4</b>																																									
<b>Neutral valence of instruction</b>	<b>Positive valence of instruction</b>	<b>Neutral valence of instruction</b>	<b>Positive valence of instruction</b>																																								
<table border="0"> <tr> <td>A</td><td>B</td><td>C</td><td>D</td></tr> <tr> <td>1 [X]</td><td>[ ]</td><td>[ ]</td><td>[ ]</td></tr> <tr> <td>2 [ ]</td><td>[X]</td><td>[X]</td><td>[ ]</td></tr> <tr> <td>3 [ ]</td><td>[X]</td><td>[X]</td><td>[ ]</td></tr> <tr> <td>4 [X]</td><td>[ ]</td><td>[ ]</td><td>[ ]</td></tr> </table>	A	B	C	D	1 [X]	[ ]	[ ]	[ ]	2 [ ]	[X]	[X]	[ ]	3 [ ]	[X]	[X]	[ ]	4 [X]	[ ]	[ ]	[ ]	<table border="0"> <tr> <td>A</td><td>B</td><td>C</td><td>D</td></tr> <tr> <td>1 [X]</td><td>[ ]</td><td>[ ]</td><td>[ ]</td></tr> <tr> <td>2 [ ]</td><td>[X]</td><td>[X]</td><td>[ ]</td></tr> <tr> <td>3 [ ]</td><td>[X]</td><td>[X]</td><td>[ ]</td></tr> <tr> <td>4 [X]</td><td>[ ]</td><td>[ ]</td><td>[ ]</td></tr> </table>	A	B	C	D	1 [X]	[ ]	[ ]	[ ]	2 [ ]	[X]	[X]	[ ]	3 [ ]	[X]	[X]	[ ]	4 [X]	[ ]	[ ]	[ ]	<p>"edit this summary"</p> <p>"The edited summary MUST NOT be between 63 and 93 words long."</p>	<p>"improve this summary"</p> <p>"The improved summary MUST NOT be between 63 and 93 words long."</p>
A	B	C	D																																								
1 [X]	[ ]	[ ]	[ ]																																								
2 [ ]	[X]	[X]	[ ]																																								
3 [ ]	[X]	[X]	[ ]																																								
4 [X]	[ ]	[ ]	[ ]																																								
A	B	C	D																																								
1 [X]	[ ]	[ ]	[ ]																																								
2 [ ]	[X]	[X]	[ ]																																								
3 [ ]	[X]	[X]	[ ]																																								
4 [X]	[ ]	[ ]	[ ]																																								
<p>"change [...] so that the grid is perfectly symmetrical"</p>		<p>"improve [...] so that the grid is perfectly symmetrical"</p>																																									

**Figure 1.** We assigned the different tasks, symmetry task and summary task, as well as the manipulations of the solution efficiency or the valence of instruction to the four experiments.



**Figure 2.** Percentage of chosen additive solution strategy depending on solution efficiency (Experiments 1 and 2) or depending on valence of instruction (Experiments 3 and 4): GPT-4 vs. humans in symmetry task (Experiments 1 and 3) or in summary task (Experiments 2 and 4).

**Table 1.** Frequency of the chosen additive or subtractive solution strategy in humans and GPT-4.

Chosen solution strategy	Agent		Total ( <i>n</i> )
	Human ( <i>n</i> )	GPT-4 ( <i>n</i> )	
Additive	335	481	816
Subtractive	253	199	452
Total	588	680	1268

## SI Appendix

**Table S1.** Exploratory analysis of the chosen solution strategy for each condition.

		Chosen additive strategies (in %)	Chosen subtractive strategies (in %)	$\chi^2$	df	p
Human	<b>Experiment 1</b>					
	Addition/subtraction equal	65.7	34.3	6.58	1	.01
	Subtraction more efficient	29.2	70.8	11.22	1	< .001
	<b>Experiment 2</b>					
	Addition/subtraction equal	57.5	42.5	1.94	1	.16
	Subtraction more efficient	39.3	60.7	3.86	1	.05
	<b>Experiment 3</b>					
	Valence neutral	63.3	36.7	4.27	1	.04
	Valence positive	61.0	39.0	2.86	1	.09
	<b>Experiment 4</b>					
	Valence neutral	65.1	34.9	7.53	1	< .01
	Valence positive	73.5	26.5	18.33	1	< .001
GPT-4	<b>Experiment 1</b>					
	Addition/subtraction equal	54.1	45.9	0.58	1	.45
	Subtraction more efficient	69.4	30.6	12.81	1	< .001
	<b>Experiment 2</b>					
	Addition/subtraction equal	77.7	22.3	25.99	1	< .001
	Subtraction more efficient	90.6	9.4	56.01	1	< .001
	<b>Experiment 3</b>					
	Valence neutral	45.9	54.1	0.58	1	.45
	Valence positive	55.3	44.7	0.95	1	.33
	<b>Experiment 4</b>					
	Valence neutral	76.5	23.5	23.82	1	< .001
	Valence positive	96.5	3.5	73.42	1	< .001

**Table S2.** Overview of accuracy of solutions proposed by humans and by GPT-4.

	Human		GPT-4	
	Accurate Solutions (n)	Inaccurate Solutions (n)	Accurate Solutions (n)	Inaccurate Solutions (n)
<b>Experiment 1</b>				
Addition/subtraction equal	65 (97.0%)	2 (3.0%)	2 (2.4%)	83 (97.6%)
Subtraction more efficient	52 (80.0%)	13 (20.0%)	2 (2.4 %)	83 (97.6%)
<b>Experiment 2</b>				
Addition/subtraction equal	74 (85.1%)	13 (14.9%)	15 (17.6%)	70 (82.4%)
Subtraction more efficient	73 (86.9%)	11 (13.1%)	13 (15.3%)	72 (84.7%)
<b>Experiment 3</b>				
Valence neutral	58 (96.7%)	2 (3.3%)	1 (1.2%)	84 (98.8%)
Valence positive	55 (93.2%)	4 (6.8%)	3 (3.5%)	82 (96.5%)
<b>Experiment 4</b>				
Valence neutral	68 (81.9%)	15 (18.1%)	15 (17.6%)	70 (82.4%)
Valence positive	72 (86.7%)	11 (13.3%)	57 (67.1%)	28 (32.9%)

**Table S3.** Detailed breakdown of human participant exclusions.

	Before exclusion	Double participation	No initial consent	No final consent	Page lost focus	Equal box/word count	Final sample
<b>Experiment 1</b>							
Addition/subtraction equal	191	0	0	1	31	92	67
Subtraction more efficient	124	1	1	3	19	35	65
Total	315	1	1	4	50	127	132
<b>Experiment 2</b>							
Addition/subtraction equal	107	0	0	0	17	3	87
Subtraction more efficient	99	0	0	1	12	2	84
Total	206	0	0	1	29	5	171
<b>Experiment 3</b>							
Valence neutral	178	0	0	1	34	83	60
Valence positive	174	0	0	1	28	86	59
Total	352	0	0	2	62	169	119
<b>Experiment 4</b>							
Valence neutral	107	0	1	0	17	6	83
Valence positive	121	0	0	0	32	6	83
Total	228	0	1	0	49	12	166

**Table S4.** Demographic characteristics of the participants for all four experiments.

	n	Gender				Age (in years)		
		Male	Female	Non-binary	Not specified	M	SD	Range
<b>Experiment 1</b>	132	63	63	6	0	34.0	10.9	18–67
<b>Experiment 2</b>	171	79	88	3	1	37.2	12.3	20–75
<b>Experiment 3</b>	119	65	50	4	0	35.1	10.8	19–68
<b>Experiment 4</b>	166	79	83	3	1	37.0	12.0	19–74

*Note.* One participant's age was excluded from Experiment 4 due to an implausible value.

**Table S5.** Detailed breakdown of GPT-4 iteration exclusions.

	Before exclusion	Incomplete answers	Equal box/word count	Final sample
<b><i>Experiment 1</i></b>				
Addition/subtraction equal	107	0	22	85
Subtraction more efficient	94	0	9	85
Total	201	0	31	170
<b><i>Experiment 2</i></b>				
Addition/subtraction equal	91	0	6	85
Subtraction more efficient	86	0	1	85
Total	177	0	7	170
<b><i>Experiment 3</i></b>				
Valence neutral	112	1	26	85
Valence positive	112	1	26	85
Total	222	2	52	170
<b><i>Experiment 4</i></b>				
Valence neutral	89	0	4	85
Valence positive	86	0	1	85
Total	175	0	5	170

**Table S6.** Median completion times across the two surveys for all four experiments.

	Median time to complete the experiment (in minutes)	
	First survey	Additional survey
<b><i>Experiment 1</i></b>	03:10	03:12
<b><i>Experiment 2</i></b>	04:53	06:06
<b><i>Experiment 3</i></b>	02:53	03:00
<b><i>Experiment 4</i></b>	05:29	07:50

## Vignettes

The following section presents detailed vignettes that illustrate specific scenarios and tasks for the symmetry task and the summary task, each with the condition of “addition and subtraction equally efficient” with neutral instructional valence. The changes include shortening the instructions and changing “toggle the colour of any box” to “switch any field from [X] to [ ] or from [ ] to [X]” to make the instructions more understandable.

**Symmetry task.** There is a digital 4-by-4 grid with columns labeled A through D from left to right and rows labeled 1 through 4 from top to bottom. You can switch any field from "[X]" to "[ ]" or from "[ ]" to "[X]". Below you can see the current pattern of fields.

A	B	C	D
1	[X]	[ ]	[ ]
2	[ ]	[X]	[X]
3	[ ]	[X]	[X]
4	[X]	[ ]	[ ]

Your task is to change the pattern of the fields so that the grid is perfectly symmetrical from left to right and from top to bottom, that is, it must be exactly mirror symmetrical both horizontally and vertically. There are many technically correct solutions to this puzzle. However, your task is to create symmetry with as little switching as possible.

Please describe how you solve the task in the following text box.

[For humans, a text box was displayed here]

[For GPT-4] Text box:

**Summary task.** Here is a short news article about fire fighting in New South Wales, Australia. Please read it carefully.

'Thor' aircraft, capable of dropping 15,000 litres of water, arrives in NSW ahead of bushfire season

(By Nick Dole, ABC News, Sep 1, 2015)

The New South Wales Rural Fire Service has unveiled its newest firefighting tool, a massive water-bombing aircraft nicknamed Thor, ahead of the bushfire season. The plane has just arrived from North America for the Australian bushfire season and performed a practise water-bombing exercise near Richmond air base in north-west Sydney. The air tanker can drop more than 15,000 litres of water or fire retardant on a blaze within a few seconds. It has a loaded cruising speed of 545 kilometres an hour, meaning it will be able to reach most fires in the state within one hour. Emergency Services Minister David Elliott said the aircraft would be a welcome sight for residents in fire-prone areas. "The people of NSW will be able to rest easy tonight, the first day of spring, knowing that Thor will be in the skies looking out for them," Mr Elliott said. RFS Commissioner Shane Fitzsimmons said the specially configured C-130 Hercules will dwarf the capability of the smaller water-bombing planes in the fleet. "This can carry five times the water of those machines," he said. "They can take off and land on airstrips of less than two kilometres, they're very versatile." The plane's pilot, Rickey Rau said Thor can fly fully loaded for at least five hours. "Depending on where the fire is, we can launch out of here fully loaded and do two to three runs on the fire before needing fuel," Mr Rau said. The plane will be based at the Richmond air base, but could also be moved to smaller bases at Williamtown, Tamworth, Dubbo and Canberra. The aircraft will be leased to NSW for two fire seasons as part of a \$10 million State Government funding package. An ever bigger DC-10 air tanker will arrive in October. Commissioner Fitzsimmons said much of NSW was facing an "above normal" risk of bushfires this season. "That's largely caused by the moisture deficit," he said. He said there has been solid rainfall in recent months, but it was not likely to have a lasting effect. "With the onset of a strengthening El Nino still dominating the forecast for this season, the moisture is expected to be depleted fairly quickly, giving rise to a difficult fire season for 2015/16," he said.

This is again the article you have just read. Below the article you will find a summary of the article.

'Thor' aircraft, capable of dropping 15,000 litres of water, arrives in NSW ahead of bushfire season

(By Nick Dole, ABC News, Sep 1, 2015)

The New South Wales Rural Fire Service has unveiled its newest firefighting tool, a massive water-bombing aircraft nicknamed Thor, ahead of the bushfire season. The plane has just arrived from North America for the Australian bushfire season and performed a practise water-bombing exercise near Richmond air base in north-west Sydney. The air tanker can drop more than 15,000 litres of water or fire retardant on a blaze within a few seconds. It has a loaded cruising speed of 545 kilometres an hour, meaning it will be able to reach most fires in the state within one hour. Emergency Services Minister David Elliott said the aircraft would be a welcome sight for residents in fire-prone areas. "The people of NSW will be able to rest easy tonight, the first day of spring, knowing that Thor will be in the skies looking out for them," Mr Elliott said. RFS Commissioner Shane Fitzsimmons said the specially configured C-130 Hercules will dwarf the capability of the smaller water-bombing planes in the fleet. "This can carry five times the water of those machines," he said. "They can take off and land on airstrips of less than two kilometres, they're very versatile." The plane's pilot, Rickey Rau said Thor can fly fully loaded for at least five hours. "Depending on where the fire is, we can launch out of here fully loaded and do two to three runs on the fire before needing fuel," Mr Rau said. The plane will be based at the Richmond air base, but could also be moved to smaller bases at Williamtown, Tamworth, Dubbo and Canberra. The aircraft will be leased to NSW for two fire seasons as part of a \$10 million State Government funding package. An ever bigger DC-10 air tanker will arrive in October. Commissioner

Fitzsimmons said much of NSW was facing an "above normal" risk of bushfires this season. "That's largely caused by the moisture deficit," he said. He said there has been solid rainfall in recent months, but it was not likely to have a lasting effect. "With the onset of a strengthening El Nino still dominating the forecast for this season, the moisture is expected to be depleted fairly quickly, giving rise to a difficult fire season for 2015/16," he said.

**SUMMARY:**

According to ABC's Nick Dole, September 1st, the New South Wales Fire Service will use the first of two fire-fighting US-made air tankers this spring. Lack of moisture and El Nino signal more risk of fire. But the plane, reaching any fire in NSW quickly and dumping thousands of litres quickly, will make people feel safer. Fuel efficiency, range, speed, loading capacity and its ability to use short airstrips make it a singularly versatile fire-fighting tool.

**YOUR TASK:** We would like you to edit this summary. For your convenience, we have pasted it into the box below.

**NOTE:** The summary currently contains 76 words. The edited summary MUST NOT be between 64 and 88 words long.

[For humans, a text box was displayed here]