HHAI 2023: Augmenting Human Intellect
P. Lukowicz et al. (Eds.)

doi:10.3233/FAIA230081

Interacting with Large Language Models: A Case Study on AI-Aided Brainstorming for Guesstimation Problems

Vildan SALIKUTLUK [a,1], Dorothea KOERT [a] and Frank JÄKEL [a]
[a] *Centre for Cognitive Science, Technical University of Darmstadt, Germany*
ORCiD ID: Vildan Salikutluk https://orcid.org/0000-0001-7913-7349, Dorothea Koert https://orcid.org/0000-0002-3571-6848, Frank Jäkel https://orcid.org/0000-0002-1355-7663

Abstract. Designing cooperative AI-systems that do not automate tasks but rather aid human cognition is challenging and requires human-centered design approaches. Here, we introduce AI-aided brainstorming for solving guesstimation problems, i.e. estimating quantities from incomplete information, as a testbed for human-AI interaction with large language models (LLMs). In a think-aloud study, we found that humans decompose guesstimation questions into sub-questions and often replace them with semantically related ones. If they fail to brainstorm related questions, they often get stuck and do not find a solution. Therefore, to support this brainstorming process, we prompted a large language model (GPT-3) with successful replacements from our think-aloud data. In follow-up studies, we tested whether the availability of this tool improves participants' answers. While the tool successfully produced human-like suggestions, participants were reluctant to use it. From our findings, we conclude that for human-AI interaction with LLMs to be successful AI-systems must complement rather than mimic a user's associations.

Keywords. Human-AI Interaction, Guesstimation, Large Language Models

1. Introduction

Recent artificial intelligence (AI) systems, in particular large language models (LLMs), show great potential to support human problem solving [1,2,3,4]. The availability of tools such as ChatGPT [5] and OpenAssistant [6] now allows the general public to use LLMs for different tasks. Nevertheless, it remains an active research question how to best design cooperative AI systems that do not fully automate a task but rather aid human problem solving. In this paper, we propose that *guesstimation problems*, i.e. problems that require estimates of unknown quantities when precise quantitative modeling is not an option [7], are a promising target for cooperative AI support with LLMs.

Guesstimation problems are omnipresent in forecasting scenarios [8]. Business consultants, intelligence analysts, political pundits, or risk assessors constantly work on guesstimation-like problems [9]. This is also why guesstimation questions are commonly

[1]Corresponding Author: Vildan Salikutluk, vildan.salikutluk@tu-darmstadt.de

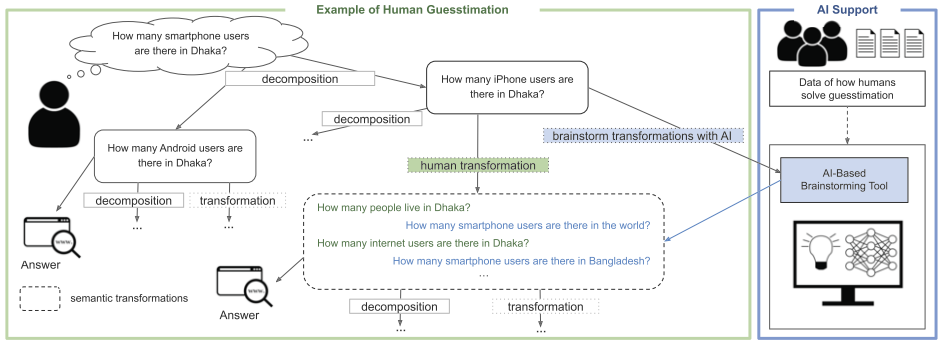


Figure 1. Example of human guesstimation with AI support. Results from our think-aloud study show that besides decomposition of questions into sub-questions also brainstorming semantic transformations of (sub-)questions is a crucial step in the solution process. We propose that when subjects get stuck in the solution process AI-based support for brainstorming more suggestions can be beneficial. Thus, our LLM-based tool, which we prompted with examples from think-aloud data, returns human-like semantically transformed questions (in blue). Subjects can then answer a transformed question directly, transform or decompose it further.

used in assessment centers, where potential employees are challenged to answer questions such as 'How many golf balls fit into a school bus?' or 'How many smartphone users are there in Dhaka?'. These questions usually require a decomposition into sub-questions (e.g. 'How many people live in Dhaka?' and 'Who uses a smartphone?'), and the ability to solve such problems can serve as an indicator of a candidate's mental flexibility and quantitative abilities [10]. Such deliberation is not only central to guesstimation but improves general problem solving skills [11]. As human performance in solving guesstimation problems varies widely [8] there is great potential for AI support.

In this paper, we investigate three research questions (RQ1 - RQ3) related to potential AI support for guesstimation. First, we investigate *how humans solve guesstimation problems and what common impasses occur during the solution process* (**RQ 1**) in a think-aloud study. We find that it is not only important to decompose guesstimation questions into good sub-questions [10] but that brainstorming semantic transformations of the (sub-)question at hand is crucial for solving guesstimation problems successfully. These findings align with previous work indicating that successful forecasters consider more, and more detailed decompositions [8,9,12], have an open mind, consider more options and information [13]. Additionally, our results show that when participants cannot brainstorm variations and generate related questions, they often get stuck and even fail to produce any answers to guesstimation questions.

While there are efforts to design tools to improve forecasting [14], there is none to support brainstorming semantic transformations in guesstimation tasks specifically. Furthermore, our goal is not automation – in contrast to some existing AI systems that aim to (semi-)automate guesstimation [15,16,17]. Instead, we follow a human-centered design approach [18] and identify potential targets for AI support.

Inspired by successful applications of natural language processing (NLP) to generate ideas or aid in various brainstorming tasks [2,19,20], writing [21,22,3], or mood board creation [2], we propose the use of an AI-aided brainstorming tool for the specific use-case of solving guesstimation problems. More specifically, we use the *Generative Pre-trained Transformer 3* (GPT-3), which is an LLM [23] that was already used successfully

in several different application areas [1,24,25,26,27,28]. We provide successful semantic transformations of (sub-)questions that we collected in our think-aloud study as examples for GPT-3 to teach it to produce similarly useful transformations.

In follow-up experiments, we evaluate *whether GPT-3 can be prompted successfully with think-aloud data to brainstorm human-like suggestions for given (sub-)questions* (**RQ 2**). Subsequently, we conduct a study in which we provide our tool to participants to test *whether the availability of our AI-based brainstorming tool affects human performance on guesstimation problems* (**RQ 3**). Fig. 1 illustrates the proposed approach.

The main contributions of this paper are the following. First, we introduce guesstimation problems as a suitable testbed for cooperative human-AI interaction with LLMs. Second, with a think-aloud study, we show at which points such a system might support humans during guesstimation, identifying brainstorming relevant (sub-)questions as an important target. Third, we use the think-aloud data to prompt an LLM, specifically GPT-3, with successful semantic transformations and show that this brainstorming tool provides human-like suggestions. Lastly, we conduct an evaluation study to test how the availability of such an AI-based brainstorming tool influences guesstimation.

2. Related Work

Guesstimation. Guesstimation problems are commonplace and the ability to solve them is often key in high-stakes decisions in politics or business [8,9]. Furthermore, it was shown that training students on solving guesstimation problems improves their general problem solving abilities and fosters important skills for STEM subjects [11,29,30].

There are AI-systems that can solve some guesstimation problems [15,17,31,32]. The Back-of-the Envelope-Solver (BotE-Solver), e.g., is a combination of a large knowledge base and quantitative strategies, which transform and decompose a given guesstimation problem into potentially easier sub-problems. It can answer a small set of test questions correctly within one order of magnitude (8 questions in [15] and 13 in [31]). Another system is GORT: Guesstimation with Ontologies and Reasoning Techniques [17,32], a semi-automated system that combines semantic web technology with planning and reasoning methods. It either answers questions directly using semantic web technology or decompose them into sub-questions. If it is unable to answer the sub-questions directly, it asks the user for one and uses it to calculate an answer. Both systems show promising results but are limited in how many and what kind of questions they can answer, e.g. because their solution strategies are domain-specific rules [32] or their database is too small [31]. Also, both approaches focus on (semi-)automated solutions, i.e. they are not aimed at providing support for humans during their solution process. While the strategies used in these systems seem psychologically plausible and might therefore be used interactively to support humans, they are not based on empirical data on how humans actually solve guesstimation problems.

Large Language Models. LLMs have already been adapted to programming tasks [24, 1] and quantitative reasoning problems [27], and along with other NLP approaches have previously been proposed to generate ideas or aid humans in various brainstorming tasks [19,20]. Even if the suggestions of a brainstorming AI assistant might not all be good, they can be tweaked by the user or spark related and better ideas. This was already shown for other example scenarios such as writing [21,22,33,3] or mood board creation [2].

There are efforts to develop tools to improve forecasting with machine learning, expert knowledge, and crowd-sourcing [14] and LLMs have recently been proposed for decompositional reasoning [34,35]. However, to the best of the authors' knowledge, there is currently no LLM-based approach to support brainstorming in guesstimation.

Interaction in Human-AI Teams. Even though many tasks require human creativity and judgment, AI systems have the potential to support humans to increase their productivity and improve the overall results of the human-AI team, e.g. with programming or writing assistants [24,1,36,37,22,3]. However, designing human-AI interaction for such scenarios is not easy [18,38], and updated guidelines for interaction and interface design are required [39]. This might also be why such hybrid teams sometimes do not improve performance [37]. Furthermore, the best possible hybrid team performance is not necessarily achieved by combining the most accurate systems with its user but rather those that are complementary [40,41] and use them as decision support [42]. Therefore, how the interaction of a hybrid team needs to be designed, how the team's performance is evaluated, and how successful they are is case-specific and achieving beneficial team synergy can be difficult [37]. Overall, how to best achieve human-AI synergy remains an important open research question.

3. AI-Aided Brainstorming to Support Humans During Guesstimation

In this section, we introduce and evaluate our approach for AI-aided brainstorming to support guesstimation. In Section 3.1, we follow a human-centered approach by investigating how humans solve guesstimation problems to identify where they could benefit from AI support. Our results indicate that a crucial step to finding good answers is re-framing (sub-)questions into semantically related ones. Thus, we propose the use of an LLM to support humans during this step in Section 3.2. In Section 3.2.1, we show that the model, which we prompted with successful transformations from our think-aloud study, produces human-like suggestions. Subsequently, in Section 3.2.2, we present an evaluation study on the effect of our LLM-based brainstorming tool on the performance of humans solving guesstimation problems. Fig. 2 presents an overview of these studies.

3.1. Understanding Impasses in Human Guesstimation

While there are studies with forecasting experts [9] or best practices and example solutions for guesstimation-like problems [7,10,43], there is a lack of empirical studies on the underlying solution process and potential impasses in human guesstimation. In this section we present a think-aloud study on human guesstimation, which indicates that a crucial step to successfully finding answers is re-framing guesstimation questions into semantically related ones.

Methods. We conducted a think-aloud study with 6 participants (3 female, 3 male, 20-24 years old, all were university students that received partial course credit for participation). The study was conducted in their native language (German). The local ethics board approved the study, and all participants provided informed consent. Each of the participants was tasked to solve 10 guesstimation problems and think aloud while doing so. The questions are chosen to cover a wide range of different domains and topics, e.g.

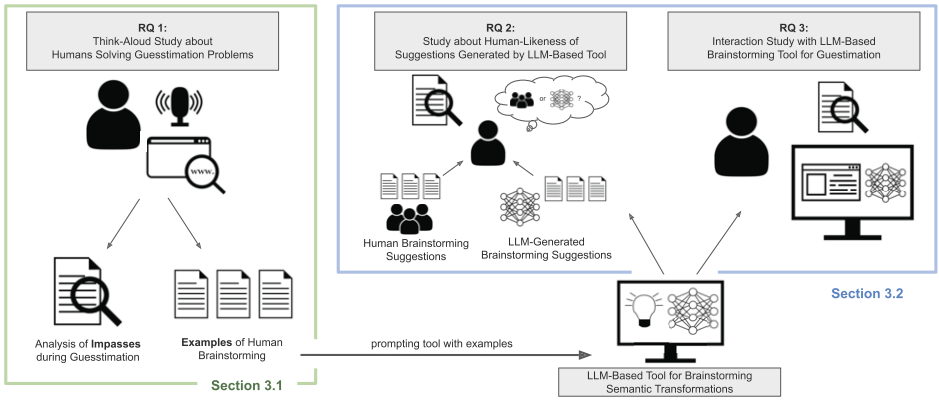


Figure 2. Overview of our studies following a human-centered approach to support humans during guesstimation with an AI-based tool. In a think-aloud study (Section 3.1), we identify brainstorming of semantic transformations of (sub-)questions as a common impasse. Therefore, we propose the use of a LLM for brainstorming by prompting it with examples from our think-aloud study. We evaluate the human-likeness of resulting suggestions in Section 3.2.1 and investigate effects of our tool on human guesstimation in Section 3.2.2.

'How many pizzas are delivered daily in Darmstadt?' or 'How many smartphones are sold per minute in Germany?'. Participants had seven minutes to come up with their best estimate per question. They were allowed to research anything they wanted on Google Search, take notes, and use a calculator. The interface for this study is shown in Fig. 3 (but the LLM-based brainstorming tool was not present). During the experiment, we recorded a screen capture video, think-aloud audio data, their search terms for Google Search as well as their notes and calculations. Since the answers were impossible to find directly through Google Search, participants had to decompose the questions into subquestions and think about different approaches to the problem. The video and audio data were transcribed and analyzed with the grounded theory approach [44,45].

Results. Most subjects were generally able to answer the guesstimation questions. Overall, we collected 60 trials. We excluded the trials where subjects stated that their answers were pure guesses or when questions remained unanswered. The analysis of the remaining 43 trials reveals different strategies the participants use to solve the guesstimation problems. A particularly important strategy for constructing reasonable answers is semantic transformation of a question into a related one. E.g., a participant was unable to find an answer to a sub-question they worked on, like 'How much does a female student weigh?'. They then replaced the question with 'How much does a woman weigh?'. Although the two questions have different answers since weight varies with age, the answer to the second question was accessible online. Thus, the participant used it to answer the original question since the two answers are not too different and therefore the error due to this substitution seemed tolerable for the final estimate.

In our data, we found three different semantic transformations. Either participants generalize a concept (e.g. Portuguese citizens to Europeans), or they specialize it (e.g. limousine to limousine of a specific brand). They also often transform a concept into a related one (Portuguese citizens to German citizens). On the left of Fig. 1 (in green) and in Table 1 such transformations are shown. Overall, we collected 15 suitable examples

Table 1. Examples of semantic transformations for guesstimation problems from the think-aloud data. Bold concepts were replaced during the transformation from the seed questions which participants worked on.

Seed Question	Transformed Question	Transformation
How much does a **female student** weigh?	How much does a **woman** weigh?	Generalization
	How much does a **student** weigh?	Generalization
How many **trains** depart from a single platform daily at the main station in Berlin?	How many **long distance trains** depart from a single platform daily at the main station in Berlin?	Specialization
	How many **express trains** depart from a single platform daily at the main station in Berlin?	Specialization
How many grams of chocolate are in a **Mars** bar?	How many grams of chocolate are in a **Twix** bar?	Related Concept
	How many grams of chocolate are in a **Bounty** bar?	Related Concept

for semantic transformations. In Section 3.2 we will use them to prompt an LLM to brainstorm relevant substitutions.

Importantly, the protocols reveal that when participants were unable to find an appropriate substitution or decomposition for a question, they were unable to answer reasonably. Of the remaining 43 trials participants were completely stuck and just guessed in 12 (at least once per participant). Participants also often indicated that their current strategy was not the best and they wished they had a better idea. Overall, this occurred 66 times across all 43 trials. Considering that subjects repeatedly got stuck in some way (78 times overall), we hypothesize that new ideas and semantically reasonable substitutions for the current (sub-)question would have been helpful to the participants.

3.2. Brainstorming with a Large Language Model for Guesstimation

Since the results of the think-aloud study show that impasses occur when humans are unable to generate semantically related questions, we hypothesize that a tool for semantic brainstorming could be beneficial during guesstimation. Specifically, we prompted the *Generative Pre-trained Transformer 3* (GPT-3) with successful semantic transformations from the think-aloud protocols. GPT-3 is an LLM pre-trained on a vast corpus of language data, such that it can be instructed to perform a new language task by prompting it with a natural language description of what it should do (the instruction) and only a few appropriate input-output example pairs (few-shot learning) [23]. While GPT-3 is able to produce novel text in response to a prompt [25], its performance for a given task strongly depends on the examples it is presented with [46].

We prompted GPT-3 with the following instructions: 'For each question about an object below, I'll suggest a helpful related question – a more general question, a more specific question, or a question with an answer I can otherwise directly relate to the original answer.' They were followed by the pairs of original and rephrased questions from the think-aloud protocols. Examples of semantic transformations from the think-aloud study used to prompt GPT-3 are shown in Table 1. Overall, we used 15 semantic transformation examples. We accessed GPT-3 through *Elicit* by Ought, Inc. [47]. The tool is included in the user interface from the think-aloud study shown in Fig. 3 (a). In

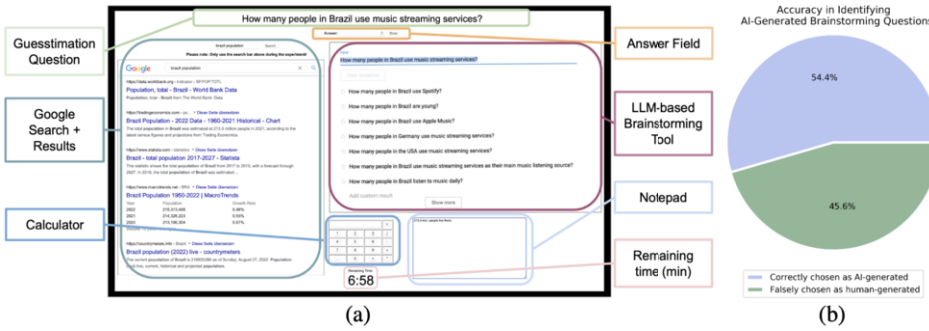


Figure 3. (a) Interface used in our guesstimation experiments. In our think-aloud study (Section 3.1), subjects could solve the presented guesstimation question by using Google, a notepad and a calculator. In our evaluation study (Section 3.2.2) they could additionally use the LLM-brainstorming tool. The tool and its suggestions for related questions for the previously unseen input question 'How many people in Brazil use music streaming services?' are shown. (b) Accuracy in identifying AI-generated brainstorming questions. We show the percentage of trials in which the AI-generated question among the two alternatives was correctly identified (blue) compared to trials where the human-generated question was falsely identified as generated by an AI (green).

Section 3.2.1, we evaluate if our resulting AI tool can produce human-like suggestions for brainstorming during guesstimation. Furthermore, in Section 3.2.2 we present the results of a user study where we provided our tool to humans during guesstimation.

3.2.1. Comparing Human Brainstorming and a Large Language Model

After we prompted GPT-3 with human example data from our think-aloud study, we tested if it produces human-like semantic transformations for given (sub-)questions. We first collected human transformations of guesstimation questions that were not part of the original GPT-3 prompt and used our brainstorming tool to generate semantic transformations for these questions. These transformations are then used as stimuli in a subsequent Turing-test-like experiment where we evaluate how well humans can distinguish whether a question was AI- or human-generated.

Methods. We collected human suggestions for semantic transformations of 10 participants (6 female, 4 male, 18-34 years old, all participants provided informed consent). Each subject was provided with the semantic transformation examples from our think-aloud study, which we also used to prompt GPT-3. Subsequently, we showed them 6 new guesstimation questions and asked them to brainstorm at least 7 semantically related questions. We clustered identical and semantically equivalent questions together, i.e. when they expressed the same question but the wording differed slightly. For each of the 6 guesstimation questions we selected the 7 human suggestions that were repeated most often. We compare these to the first 7 questions generated by GPT-3 (parameters: `top_p` = 0.95, `temperature` = 1 (default), `frequency_penalty` = 0.4 and duplicates were removed). We removed any typos from the human suggestions since we did not want them to be a trivial cue to distinguish human from AI suggestions. Example transformations from humans and GPT-3 are shown in Table 2.

We then asked 23 subjects (13 male, 10 female, aged between 18 and 34, all provided informed consent) to complete a two-alternative-forced-choice task (2AFC) in which they must choose which of two presented semantic transformations was AI-generated.

Table 2. Examples of most-repeated brainstorming suggestions of human subjects and GPT-3 generated suggestions for the question 'How many people use music streaming services in Brazil?'. The bold suggestions are those that were identically generated by both humans and GPT-3.

Human Suggestions (no. of subjects)	GPT-3 Questions
How many people live in Brazil? (8/10)	**How many people in Brazil use Spotify?**
How many people use Apple Music in Brazil? (6/10)	**How many people in Brazil use Apple Music?**
How many people in Brazil use Spotify? (6/10)	How many people in Brazil are young?
How many people in Brazil use Deezer? (2/10)	How many people in Germany use music streaming services?
What does a music streaming service cost in Brazil? (2/10)	How many people in the USA use music streaming services?
How many people in Brazil are listening to music? (2/10)	How many people in Brazil use music streaming services as their main music listening source?
How many people have access to the internet in Brazil? (2/10)	How many people in Brazil listen to music daily?

The study was conducted online and before the subjects started the 2AFC task, we asked them to brainstorm their own ideas for each question. This ensured that they understood what kind of suggestions they would be presented with. For each question and subject we randomly generated 7 pairs of semantic transformations from the human and AI-generated transformations (42 trials for each subject). We randomized the order in which the guesstimation questions were presented to each participant to avoid order effects.

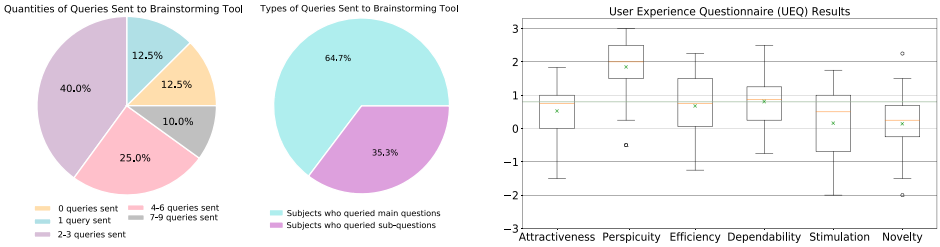
Results. We evaluate how often subjects correctly distinguished between the human and AI-generated question in the presented pairs of semantic transformations. Overall, we collected 161 trials per question (23 participants times 7 trials). Participants could not identify the AI-generated question reliably. They were unable to select the AI-generated question in 45.4% of all cases (966 trials, see Fig. 3 (b)). For two questions, the accuracy of distinguishing between human and AI-generated suggestions was even below or close to chance level (which is at 50% for 2AFC). Even though participants identified AI-generated suggestions with statistical significance ($p = 0.0029$), i.e. in 54.45% of the trials (95% CI [496 (= 51.3%), 556 (= 57.56%)]), the effect is small demonstrating that distinguishing between the human and AI-generated semantic transformation is difficult.

These results indicate that our LLM-based tool successfully produces human-like suggestions for semantic transformations. This is also confirmed by subjects' comments at the end of the experiment, e.g. 'It was very hard to guess which question was AI-generated.' and 'Sometimes I had the notion that both questions were from humans.'

3.2.2. Interactive Brainstorming with a Large Language Model

The LLM-based brainstorming tool is able to produce human-like semantic transformations of (sub-)questions. The insights from our think-aloud study (Section 3.1) and findings from [8,9,12] indicate that the availability of such a tool might be beneficial during guesstimation. Thus, we conduct a study to test what kind of effects we can observe when providing our AI-brainstorming tool while humans solve guesstimation problems.

Methods. We conducted an online study with 41 participants (23 female, 18 male, 18-39 years old). One subject had to be excluded since they did not finish the experiment. We planned the experiment for 40 participants because a power analysis indicated that



(a) (b) (c)

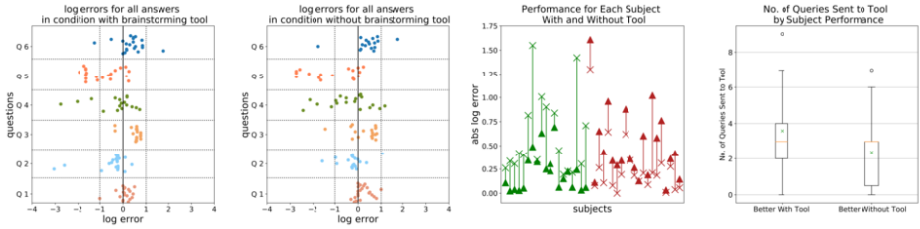
Figure 4. a) How often subjects queried our tool to brainstorm semantic transformations for their input question. (b) How many of the subjects' input questions were the main or sub-questions. (c) UEQ ratings. Values in the range of -0.8 and 0.8 are neutral, ones below -0.8 are negative and ones greater than 0.8 are positive (marked as green line). Green crosses show the mean and orange lines median values.

a paired t-test could find a medium-sized effect of the tool with high probability. The study was conducted in English and was approved by the local ethics board. All participants provided informed consent before the study started. The online study started with a short video call for on-boarding and setting up. The instructions were explained and a test trial (with the brainstorming tool) was completed. After the call ended, participants answered each of the 6 guesstimation questions within eight minutes. The questions were the same as the ones used in Section 3.2.1. The subjects used the same interface as in the think-aloud study (Fig. 3 (a)), where we additionally included the AI-brainstorming tool. Their final answers, as well as their notes, calculations and, Google search terms were collected. Furthermore, everything they typed into our brainstorming tool as well as the tool's suggestions based on their input were recorded.

We knew the correct answers for the guesstimation questions in this study, e.g. through paid services like statista.com, to compute the accuracy of the subjects' estimates. However, the participants could not access these paid services and we checked that the answers could not be found directly through Google. We used a within-subject design where each participant completed 2 blocks of 3 questions each, one of which they solved with access to the brainstorming tool and the other one without. We counterbalanced the order of the question blocks. Which block was answered with or without the tool and whether subjects started a question block with the tool or not was counterbalanced as well. Within the blocks, we randomized the questions for each participant.

After the block with the brainstorming tool, subjects completed the User Experience Questionnaire (UEQ) [48] about our brainstorming tool. The UEQ measures how users evaluate pragmatic qualities (efficiency, perspicuity, dependability) and hedonic qualities (originality and stimulation). After each block participants also rated on a 5-point Likert scale if they knew how to approach the questions to get the best possible answer (S1), if they thought their answer was good (S2), and if they wished for more tools to help with the task (S3). Lastly, they had the option to report comments.

Results. We analyze how often participants used the brainstorming tool. Overall, 35 out of 40 participants queried the tool during the experiment. Fig. 4 (a) summarizes the number of queries per subject. Furthermore, we analyze what types of questions subjects brainstormed with the tool. The participants often used the main question as input to the tool. 35.3% of subjects also brainstormed sub-questions with the tool (see Fig. 4 (b)). These sub-questions were used to brainstorm more specific aspects of the main questions.



(a) (b) (c) (d)

Figure 5. (a) $\log_{10}$ ratios of the responses to the true values, i.e. the deviation of subjects' answers for each question when working with the brainstorming tool. (b) $\log_{10}$ ratios when they worked without our tool. (c) performance changes of each subject in the condition with and without our tool. Green lines indicate subjects with better performance when they worked with the brainstorming tool. Red lines represent subjects with lower performance with the tool. The crosses show the performance without the tool and triangles show the performance with the tool. (d) how often our tool was used by subjects who performed better with the tool and those whose performance was lower in the same condition (green cross = mean; orange line = median).

We also analyzed how subjects perceive the tool by evaluating their answers to the UEQ. Answers to the UEQ are positive if their value is above 0.8, neutral between -0.8 and 0.8, and negative if they are below -0.8. Our results show a positive evaluation for perspicuity (mean score = 1.85, SD = 0.68), meaning that the tool is e.g. understandable. A positive evaluation was given for dependability (mean score = 0.81, SD = 0.57), meaning that the tool is e.g. supportive. Further, the items about how motivating (mean = 0.8, SD = 1.1) and good (mean = 0.8, SD = 1.0) the tool is, are almost rated positive. All other items were neutral. Results of the UEQ constructs are summarized in Fig. 4 (c).

In both conditions, we collected 120 trials overall (40 participants, 3 questions each). We excluded trials in which no answers were provided (6 with the tool and 8 without it). We evaluate the influence of our tool on the subjects' performance in solving guesstimation problems. We define the response error as the $\log_{10}$ ratio of the given response of the participants to the true value. A perfect response has a value of 0, and a value greater than 1 or smaller than -1 means that the participant was off by a factor of ten, i.e. one order of magnitude. The errors of all responses sorted by question can be seen in Fig. 5 (a) for the condition with the tool and in Fig. 5 (b) without it. Participants were able to answer the guesstimation questions with being less than one order of magnitude off for most questions. We also evaluate performance of each subject individually (with the absolute log ratio for each condition). Overall, 19 subjects had better accuracy in the condition with the tool (see green lines in Fig. 5 (c)). Fig. 5 (d) compares the number of queries from the subjects whose performance was higher in the condition with the tool to the number of queries of subjects whose performance was lower. Subjects with better performance in the condition with the tool used it more often on average (3.6 times) than those with lower performance (2.4 times), but the difference is not significant ($p = 0.088$). Furthermore, we compute the mean absolute response error over the three questions for each subject in each of the two conditions and put these measures into a dependent t-test ($p = 0.32$) as well as a Wilcoxon signed rank test ($p = 0.31$). Neither revealed a significant difference in the quality of answers. All response errors for each question are shown in Fig. 5 (a) and (b). Neither test revealed any significance regarding the response times either (t-test $p = 0.42$; Wilcoxon $p = 0.43$). Lastly, evaluating the scores of the statements (S1 - S3) reveals a significant difference for S3 (t-test $p = 0.042$), i.e. 'I wish I had more

tools and help during the task', for participants who started with the brainstorming tool and answered the second half of the questions without it.

Overall, participants rated the tool positively in the UEQ, and some even commented that it is 'cool' and 'helpful'. Also, subjects who started the trials with the brainstorming tool indicated that they wished for more help/tools when they had to answer the remaining questions without it (S3). However, we did not see a significant effect of the tool's availability on the participants' performance. Some comments of the subjects reveal reasons for not using the tool, e.g. its suggestions being similar to those that Google presents as 'people also ask' or that its suggestions were the same they had in mind already.

4. Discussion and Conclusion

Due to the fast progress in AI some tasks that are currently performed by humans might be automated soon [26,49,50]. However, when human creativity and judgment are required, AI systems will support humans and increase their productivity but are unlikely to replace them completely. Good examples for this are programming or writing assistants [24,1,36,37,22,3]. When there is no clear division of labor and full automation is not the aim human-AI interaction remains challenging [18,38], and guidelines are needed [39]. Thus, human-centered design approaches and identifying suitable testbeds are key for developing AI systems that can support human problem solving [38].

Guesstimation as a Testbed for Human-Centered AI. Conceiving of scenarios in which AI systems can support humans despite their current constraints is not trivial [22]. Here, we propose guesstimation problems as an interesting testbed for human-AI interaction with LLMs. Not only can such problems be studied in the lab and performance can be scored quantitatively (cf. Fig. 5 (a) and (b)), they are also not just toy problems: Guesstimation is challenging for both humans and AI systems [51] and has many real-world applications, e.g., forecasting in business and politics.

In our think-aloud study, we contribute to a better understanding of human impasses during guesstimation (**RQ1**). Specifically, it is important to consider what humans are already good at and where they can benefit from AI-support. Our results show that brainstorming semantically related (sub-)questions is central in successfully generating answers for guesstimation tasks. Hence, we present an AI brainstorming tool that can produce human-like suggestions during guesstimation (**RQ2**). Inspired by previous work that showed that humans can improve their performance when solving guesstimation-like problems by brainstorming together with other humans [9,12], we tested how brainstorming with our LLM-based tool influences human guesstimation (**RQ3**). Overall, we advocate for guesstimation as a promising application area for LLMs since it has great relevance for forecasting experts and guesstimation problems are measurable approximations of and transferable to general deliberative judgements tasks. Further, in contrast to other cooperative tasks with LLM-based systems, e.g. writing with AI-support [3,33], performance in guesstimation tasks can be objectively and quantitatively evaluated.

Limitations. Our brainstorming tool was overall perceived positively. However, the subjects in our study varied a lot in how much and in what way they used it (see Fig. 4 (a) and (b)). More than half of the subjects queried the tool only with the main question and did not use it continuously during their entire solution process. Moreover, in

our last study, the tool did not show significant effects on performance. This finding is consistent with other work on LLMs [36,37] that also showed that improving performance synergistically can be difficult in various human-AI interaction scenarios [37]. Although GPT-3 produces impressively reasonable results in our studies, subjects remarked that the tool often made suggestions they already thought of. This, again, demonstrates that LLMs capture human semantic associations well. However, the usefulness of LLM-generated suggestions for interactive brainstorming will depend on their ability to not only reproduce human-like suggestions and obvious ideas but rather complement the user's thoughts and abilities. These results also align well with other work [40,41] that shows that the best possible interaction and performance are achieved when complementary strengths of humans and AI systems are utilized. Hence, we believe the main limitation of our study has been that the tool was not specifically designed yet to complement human performance. As a lot of work on LLMs also mainly aims at imitating human performance, we consider this insight from our study an important and transferable lesson learnt for future human-AI interaction. Generally, a limitation of pre-trained models like GPT-3 are their inherent biases [23,52]. Depending on the topic of the input, it produces suggestions that perpetuate harmful stereotypes. While some research already focuses on these issues [53], if such tools are to be used for guesstimation-like problem solving in the real world, this severe issue must be addressed.

Future Directions. We propose the following future directions for semantic brainstorming during guesstimation with LLM-based tools. First, testing our proposed approach with more varied questions will be important. So far, the questions were chosen to ensure that unambiguous answers were available but cannot be googled by our participants. Thus, they had to tackle them with common-sense knowledge. Ideally, for a proper evaluation a large set of realistic forecasting questions and a comparison to expert judgments [8] is required. Specifically, the difficulty of guesstimation problems should be varied more systematically. We expect that for more difficult questions brainstorming is also harder and thus the usefulness of a AI tool might increase. Second, our prompt consisted of instructions and 15 human examples. Hence, more examples, possibly from guesstimation experts, could be added continuously such that the tool can constantly learn from human successes and improve. How prompts for LLMs are engineered can influence their outputs a lot [54]. Thus, optimizing our instruction prompt might lead to more helpful suggestions. Third, as progress on LLMs is rapid there are promising developments which could be incorporated into our tool. E.g., there is recent work on how to produce the most informative rather than the most probable output [55]. Moreover, recent work using LLMs for compositional and quantitative reasoning [34,56,35,27] could be combined with our approach to assist users with both brainstorming and with decomposing the problem. Generally, there is progress on complementing human abilities with AI, e.g. in image classification [41] or classroom settings [57] but achieving human-AI complementarity remains an open challenge. We believe that guesstimation problems provide a promising testbed to further investigate successful human-AI interaction.

Acknowledgements: The authors thank Alexandra Kraft, Adrian Kühn and Thabo Matthies for their help in parts of the experimental setup and data acquisition as well as Andreas Stuhlmüller and the Ought team for their help and access to Elicit. This work was funded by German Federal Ministry of Education and Research (project IKIDA, 01IS20045).

References

[1] Li Y, Choi D, Chung J, Kushman N, Schrittwieser J, Leblond R, et al.. Competition-Level Code Generation with AlphaCode. arXiv; 2022. Available from: https://arxiv.org/abs/2203.07814.

[2] Koch J, Lucero A, Hegemann L, Oulasvirta A. May AI? Design ideation with cooperative contextual bandits. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY: ACM; 2019. p. 1-12.

[3] Mirowski P, Mathewson KW, Pittman J, Evans R. Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems; 2023. p. 1-34.

[4] Anantrasirichai N, Bull D. Artificial intelligence in the creative industries: a review. Artificial intelligence review. 2022:1-68.

[5] OpenAI. OpenAI, editor. ChatGPT: Optimizing Language Models for Dialogue. OpenAI; 2022. Available from: https://openai.com/blog/chatgpt/.

[6] Köpf A, Kilcher Y, von Rütte D, Anagnostidis S, Tam ZR, Stevens K, et al.. OpenAssistant Conversations – Democratizing Large Language Model Alignment. arXiv; 2023. Available from: https://arxiv.org/abs/2304.07327.

[7] Weinstein L, Adam JA. Guesstimation: Solving the world's problems on the back of a cocktail napkin. Princeton, NJ: Princeton University Press; 2008.

[8] Tetlock PE, Gardner D. Superforecasting: The art and science of prediction. New York, NY: Crown Publishers; 2015.

[9] Mellers B, Stone E, Atanasov P, Rohrbaugh N, Metz SE, Ungar L, et al. The psychology of intelligence analysis: drivers of prediction accuracy in world politics. Journal of experimental psychology: applied. 2015;21(1):1.

[10] Weinstein L. Guesstimation 2.0. Princeton, NJ: Princeton University Press; 2012.

[11] Ärlebäck JB, Albarracín L. The use and potential of Fermi problems in the STEM disciplines to support the development of twenty-first century competencies. ZDM - Mathematics Education. 2019 11;51:979-90.

[12] Mellers B, Stone E, Murray T, Minster A, Rohrbaugh N, Bishop M, et al. Identifying and cultivating superforecasters as a method of improving probabilistic predictions. Perspectives on Psychological Science. 2015;10(3):267-81.

[13] Haran U, Ritov I, Mellers BA. The role of actively open-minded thinking in information acquisition, accuracy, and calibration. Judgment and Decision Making. 2013;8(3):188–201.

[14] Vaughan JW. Making Better Use of the Crowd: How Crowdsourcing Can Advance Machine Learning Research. J Mach Learn Res. 2017;18(1):7026-71.

[15] Paritosh PK, Forbus KD. Using strategies and AND/OR decomposition for back of the envelope reasoning. In: Proceedings of the 18th International Workshop on Qualitative Reasoning. Evanston, IL: Northwestern University; 2004. p. 1-7.

[16] Bundy A, Sasnauskas G, Chan M. Solving guesstimation problems using the Semantic Web: Four lessons from an application. Semantic Web. 2015;6(2):197-210.

[17] Abourbih JA, Bundy A, McNeill F. Using linked data for semi-automatic guesstimation. In: Halpin H, Chaudhri VK, Brickley D, McGuinness D, editors. Proceedings of AAAI Spring Symposium Series: Linked Data Meets Artificial Intelligence. Palo Alto, CA: AAAI Press; 2010. p. 2-7.

[18] Xu W. Toward human-centered AI: a perspective from human-computer interaction. Interactions. 2019;26(4):42-6.

[19] Li Q, Vasarhelyi M, et al. Developing a cognitive assistant for the audit plan brainstorming session. The International Journal of Digital Accounting Research. 2018;18(January):119-40.

[20] Özbal G, Pighin D, Strapparava C. Brainsup: Brainstorming support for creative sentence generation. In: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Sofia, Bulgaria: ACL; 2013. p. 1446-55.

[21] Elkins K, Chun J. Can GPT-3 pass a writer's Turing Test? Journal of Cultural Analytics. 2020;5(2):17212.

[22] Yang Q, Cranshaw J, Amershi S, Iqbal ST, Teevan J. Sketching NLP: A case study of exploring the right things to design with language intelligence. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. New York, NY: ACM; 2019. p. 1-12.

[23] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al.. Language models are few-shot learners. arXiv; 2020. Available from: https://arxiv.org/abs/2005.14165.

[24] Chen M, Tworek J, Jun H, Yuan Q, Pinto HPdO, Kaplan J, et al.. Evaluating Large Language Models Trained on Code. arXiv; 2021. Available from: https://arxiv.org/abs/2107.03374.
[25] Dale R. GPT-3: What's it good for? Natural Language Engineering. 2021 1;27:113-8.
[26] Floridi L, Chiriatti M. GPT-3: Its nature, scope, limits, and consequences. Minds and Machines. 2020 12;30:681-94.
[27] Lewkowycz A, Andreassen A, Dohan D, Dyer E, Michalewski H, Ramasesh V, et al.. Solving Quantitative Reasoning Problems with Language Models. arXiv; 2022. Available from: https://arxiv.org/abs/2206.14858.
[28] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright CL, Mishkin P, et al.. Training language models to follow instructions with human feedback. arXiv; 2022. Available from: https://arxiv.org/abs/2203.02155.
[29] Albarracín L, Gorgorió N. Devising a plan to solve Fermi problems involving large numbers. Educational Studies in Mathematics. 2014;86:79-96.
[30] Albarracín L, Gorgorió N. A brief guide to modelling in secondary school: Estimating big numbers. Teaching Mathematics and its Applications. 2015 12;34:223-8.
[31] Paritosh PK, Forbus KD. Analysis of strategic knowledge in back of the envelope reasoning. In: Proceedings of the 20th National Conference on Artificial Intelligence (AAAI-05). Palo Alto, CA: AAAI; 2005. p. 651-6.
[32] Abourbih JA. Method and system for semi-automatic guesstimation. Edinburgh; 2009.
[33] Yuan A, Coenen A, Reif E, Ippolito D. Wordcraft: Story Writing With Large Language Models. In: 27th International Conference on Intelligent User Interfaces. New York, NY: ACM; 2022. p. 841-52.
[34] Press O, Zhang M, Min S, Schmidt L, Smith NA, Lewis M. Measuring and Narrowing the Compositionality Gap in Language Models. arXiv; 2022. Available from: https://arxiv.org/abs/2210.03350.
[35] Reppert J, Rachbach B, George C, Byun LSJ, Appleton M, Stuhlmüller A. Iterated Decomposition: Improving Science Q&A by Supervising Reasoning Processes. arXiv; 2023. Available from: https://arxiv.org/abs/2301.01751.
[36] Vaithilingam P, Zhang T, Glassman EL. Expectation vs. Experience: Evaluating the Usability of Code Generation Tools Powered by Large Language Models. In: CHI Conference on Human Factors in Computing Systems Extended Abstracts. New York, NY: ACM; 2022. p. 1-7.
[37] Campero A, Vaccaro M, Song J, Wen H, Almaatouq A, Malone TW. A Test for Evaluating Performance in Human-Computer Systems. arXiv; 2022. Available from: https://arxiv.org/abs/2206.12390.
[38] Xu W, Dainoff MJ, Ge L, Gao Z. Transitioning to human interaction with AI systems: New challenges and opportunities for HCI professionals to enable human-centered AI. arXiv; 2021. Available from: https://arxiv.org/abs/2105.05424.
[39] Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, et al. Guidelines for human-AI interaction. In: CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019). New York, NY: ACM; 2019. p. 1-13.
[40] Wilder B, Horvitz E, Kamar E. Learning to complement humans. arXiv; 2020. Available from: https://arxiv.org/abs/2005.00582.
[41] Steyvers M, Tejeda H, Kerrigan G, Smyth P. Bayesian modeling of human–AI complementarity. Proceedings of the National Academy of Sciences. 2022;119(11):e2111547119.
[42] Bansal G, Nushi B, Kamar E, Horvitz E, Weld DS. Is the most accurate ai the best teammate? optimizing ai for teamwork. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35; 2021. p. 11405-14.
[43] Swartz C. Back-of-the-envelope Physics. Baltimore, MD: JHU Press; 2003.
[44] Heath H, Cowley S. Developing a grounded theory approach: a comparison of Glaser and Strauss. International journal of nursing studies. 2004;41(2):141-50.
[45] Chun Tie Y, Birks M, Francis K. Grounded theory research: A design framework for novice researchers. SAGE open medicine. 2019;7:2050312118822927.
[46] Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What makes good in-context examples for GPT-3?. arXiv; 2021. Available from: https://arxiv.org/abs/2101.06804.
[47] Ought, Inc . Elicit: The AI Research Assistant; 2022. Available from: https://elicit.org.
[48] Schrepp M, Hinderks A, Thomaschewski J. Construction of a benchmark for the User Experience Questionnaire (UEQ). International Journal of Interactive Multimedia and Artificial Intelligence. 2017;4(4):40-4.
[49] Grace K, Salvatier J, Dafoe A, Zhang B, Evans O. When will AI exceed human performance? Evidence

from AI experts. Journal of Artificial Intelligence Research. 2018;62:729-54.
[50] Matheson R. Automating artificial intelligence for medical decision-making; 2019.
[51] Evans O, Stuhlmüller A, Cundy C, Carey R, Kenton Z, McGrath T, et al. Predicting human deliberative judgments with machine learning. Oxford: Future of Humanity Institute; 2018. FHI 018-2. Available from: https://www.fhi.ox.ac.uk/wp-content/uploads/predicting-judgments-tr2018.pdf.
[52] McGuffie K, Newhouse A. The radicalization risks of GPT-3 and advanced neural language models. arXiv; 2020. Available from: https://arxiv.org/abs/2009.06807.
[53] Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A. A Survey on Bias and Fairness in Machine Learning. arXiv; 2019. Available from: https://arxiv.org/abs/1908.09635.
[54] Liu V, Chilton LB. Design guidelines for prompt engineering text-to-image generative models. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems; 2022. p. 1-23.
[55] Meister C, Pimentel T, Wiher G, Cotterell R. Typical Decoding for Natural Language Generation. arXiv; 2022. Available from: https://arxiv.org/abs/2202.00666.
[56] Stuhlmüller A, Stebbing L, Reppert J. Ought, Inc , editor. Factored Cognition Primer: How to write compositional language model programs; 2022. Accessed: Oct 12, 2022. Available from: https://primer.ought.org/.
[57] Holstein K, Aleven V. Designing for human-AI complementarity in K-12 education. arXiv; 2021. Available from: https://arxiv.org/abs/2104.01266.