

## Idea evaluation for solutions to specialized problems: leveraging the potential of crowds and large language models

Henner Gimpel, Robert Laubacher, Fabian Probst, Ricarda Schäfer, Manfred Schoch

### Angaben zur Veröffentlichung / Publication details:

Gimpel, Henner, Robert Laubacher, Fabian Probst, Ricarda Schäfer, and Manfred Schoch. 2025. "Idea evaluation for solutions to specialized problems: leveraging the potential of crowds and large language models." *Group Decision and Negotiation* 34 (4): 903–32.  
<https://doi.org/10.1007/s10726-025-09935-y>.



# Idea Evaluation for Solutions to Specialized Problems: Leveraging the Potential of Crowds and Large Language Models

Henner Gimpel<sup>1,2,3</sup> · Robert Laubacher<sup>5</sup> · Fabian Probst<sup>1,2,3</sup> · Ricarda Schäfer<sup>1,4</sup> · Manfred Schoch<sup>1,3,6</sup>

Accepted: 21 May 2025 / Published online: 28 June 2025  
© The Author(s) 2025

## Abstract

Complex problems such as climate change pose severe challenges to societies worldwide. To overcome these challenges, digital innovation contests have emerged as a promising tool for idea generation. However, assessing idea quality in innovation contests is becoming increasingly problematic in domains where specialized knowledge is needed. Traditionally, expert juries are responsible for idea evaluation in such contests. However, experts are a substantial bottleneck as they are often scarce and expensive. To assess whether expert juries could be replaced, we consider two approaches. We leverage crowdsourcing and a Large Language Model (LLM) to evaluate ideas, two approaches that are similar in terms of the aggregation of collective knowledge and could therefore be close to expert knowledge. We compare expert jury evaluations from innovation contests on climate change with crowd-sourced and LLM's evaluations and assess performance differences. Results indicate that crowds and LLMs have the ability to evaluate ideas in the complex problem domain while contest specialization—the degree to which a contest relates to a knowledge-intensive domain rather than a broad field of interest—is an inhibitor of crowd evaluation performance but does not influence the evaluation performance of LLMs. Our contribution lies with demonstrating that crowds and LLMs (as opposed to traditional expert juries) are suitable for idea evaluation and allows innovation contest operators to integrate the knowledge of crowds and LLMs to reduce the resource bottleneck of expert juries.

**Keywords** Idea evaluation · Crowdsourcing · Large language model · Specialized knowledge

## 1 Introduction

Complex problems such as climate change or social inequality are some of the most pressing challenges of our time. Complex problems share common characteristics: they are (from the current point of view) unique, and humankind has little experience with solving them (Rittel and Webber 1973). Even if solutions are implemented, it often takes years or decades until the solution quality becomes visible and measurable (Gimpel et al. 2020). Thus, complex problems pose great questions to humanity and require sophisticated problem-solving abilities, in particular, involving a broad range of affected stakeholders (Head 2008).

Online innovation contests are one tool used to foster innovative problem-solving by offering a platform that brings together people with different knowledge (Chesbrough 2006). Such platforms often address complex problems and can contain contests for different sub-problems. These sub-problems are different in their level of specialization and thus require different levels of expertise to solve them. We define the term specialization as the degree to which a contest's topic belongs to one particular, knowledge-intensive domain rather than a broad field of interest.

The ease of accessing online innovation contests usually leads to an abundance of submitted ideas, resulting in the challenge of idea evaluation (Blohm et al. 2013). There is vast agreement that groups of expert juries are the most suitable decision makers in the absence of an objectively known solution quality (Klein and Garcia 2015; Blohm et al. 2016; Görzen and Kundisch 2016; Nagar et al. 2016). In this context, decision maker refers to the instance that evaluates the idea. However, the resource of expert juries is scarce and expensive, creating a severe trade-off between efficiency and quality of evaluations. This bottleneck is one of the central challenges of using innovation contests to solve complex problems (Nagar et al. 2016).

In this paper, we therefore focus on the part of the decision-making process that deals with the evaluation of the ideas. Specifically, we study two promising alternatives to expert juries in the evaluation process, namely evaluation by crowds and Large Language Models (LLMs). These two approaches have been studied together since the emergence of LLMs, as both are similar in terms of the aggregation of collective knowledge. Crowdfunding for idea evaluation is an approach that has gained recognition from researchers and practitioners (Oosterman et al. 2014; Klein and Garcia 2015; Görzen and Kundisch 2016; Magnusson et al. 2016; Mollick and Nanda 2016; Wimbauer et al. 2019). Crowds refer to the general public (or selected communities thereof) but do not require participants to possess specific skills or expertise, in contrast to expert juries. The term “wisdom of the crowd” describes the phenomenon where the average of seemingly uninformed individual opinions can lead to collective intelligence when they work together as a group (Mollick and Nanda 2016). Existing research compares crowd and expert idea evaluations in mainly corporate contexts and finds mixed results (Magnusson et al. 2016; Wimbauer et al.

2019). Yet, most corporate idea evaluations do not belong to the class of complex problems for two reasons: (1) idea quality can often be determined quickly in corporate environments and (2) specialized knowledge is often not required as the specialization of corporate problems typically is low (Magnusson et al. 2016; Wimbauer et al. 2019). So far, research has not investigated the potential of crowdsourced idea evaluations in contexts with intransparent solution quality and varying specialization.

Coming to the second approach, the evaluation task can be automated entirely by artificial intelligence (AI), for example, machine learning models (Nagar et al. 2016). For instance, Nagar et al. (2016) offer a computational approach where a machine learning classifier analyzes different characteristics of texts to assess ideas. However, the decisions of these models are based on various parameters (such as the length of the text) and are not based on knowledge about the content of the texts to be evaluated. Currently, a promising AI-based approach for evaluation are LLMs that gained significant recognition with the release of ChatGPT at the end of 2022, particularly given that they are capable of generating output of a higher quality than previous machine learning models in certain tasks (Gao et al. 2023; Wang et al. 2023). Initial research on the capabilities of these models focuses on simple evaluation tasks such as stance detection or emotion recognition (Kocoń et al. 2023; Zhang et al. 2023). Some of these studies show that LLMs also have the ability to execute various tasks for which they are not initially trained (Kocmi and Federmann 2023). Due to this ability, LLMs may be promising for idea evaluation. However, previous research on LLMs' capabilities incorporates neither complex problems nor specialized knowledge in detail.

For applying crowdsourcing or LLMs in innovation contests for complex problems, assessing the impact of specialization on idea evaluation quality is urgently needed, as complex problems are specialized by definition, and specialization could push decision makers to their limits as they are unfamiliar with such topics. We need to understand better which decision maker can be used best in which setting in the presence of specialization. Therefore, we address the following research question:

What is the potential of crowdsourcing and LLMs for evaluating solution ideas for complex problems in the presence of specialization?

To answer this question, we compared crowd and LLM evaluations of 104 ideas from four treatment groups with different levels of specialization hosted by the platform MIT Climate CoLab with expert evaluations. Our results show that depending on the specialization crowds and LLMs have the ability to evaluate ideas in the complex problem domain. Contest specialization is a key inhibitor of crowd performance that we can counteract by adapting the evaluation procedure. In contrast, we find that LLMs can better handle different degrees of specialization and yield higher performance when using ranking tasks. For practice, we recommend that for replacing expert juries employing LLMs rather than a crowd might be appropriate.

## 2 Theoretical Background

### 2.1 Evaluating Solution Ideas for Complex Problems

Complex problems such as climate change, often also referred to as “wicked problems” are described as “complex, involving multiple possible causes and internal dynamics that could not [be] assumed to be linear, and have very negative consequences for society if not addressed properly” (Peters 2017). There is no clear path to solving these problems. However, a few principles guide the continuing efforts: (1) the need for coordination across locations, (2) the need to involve stakeholders from all types of interest groups, and (3) the need for a high degree of specialized knowledge (Stern 2006; Head 2008; Karvonen and Brand 2009). Complex problems can be addressed through ideas and innovation—the central factor for solving problems (von Hippel 1994).

Solving complex problems is a decision-making problem as it requires both the generation of multiple alternative ideas and the selection of the best solution under uncertainty and limited resources (Simon 1960). The conceptualization of the decision-making process developed by Simon (1960) divides it into three phases: intelligence, design and choice. First, the problem is analyzed, and relevant information is gathered (‘intelligence’). Then potential solutions are developed (‘design’) and finally the best option or the best options are selected (‘choice’).

To facilitate idea generation, which encompasses both the systematic analysis of the problem (‘intelligence’) and the creative development of solutions (‘design’), the concept of IT-enabled idea generation through open innovation contests has emerged (Han et al. 2020). These contests provide a structured approach to collect and refine ideas from diverse contributors. With the often abundant ideas generated in innovation contests, the challenge of effectively evaluating and processing them arises (‘choice’) (Blohm et al. 2013; Özyaygen and Balagué 2018). This challenge becomes even more demanding when considering the characteristics of complex problems. In order to evaluate solution ideas for complex problems successfully, decision makers rely on their own experience and knowledge to make a qualified decision (Gerlach et al. 2019). Making this decision in highly specialized areas requires deep knowledge of specific domains (Fischer et al. 2012).

To approach idea evaluation in the light of complex problems, we refer to knowledge regarding idea evaluation in general. Researchers studied the evaluation procedure, with absolute rating and relative ranking being two of the best-known approaches. In absolute rating, the decision maker rates alternatives independently of each other, for example, on a Likert scale. In contrast, relative ranking describes decision makers’ direct hierarchical arrangement of alternatives (Ovadia 2004). Besides the evaluation procedures, researchers studied the different types of decision makers (e.g., expert juries, external crowds, and various technical solutions) (Klein and Garcia 2015; Blohm et al. 2016; Nagar et al. 2016). Prior research suggests that expert juries generally outperform other

decision makers and, thus, have been considered closest to the “ground truth” (Klein and Garcia 2015; Blohm et al. 2016; Nagar et al. 2016). In contrast, crowds and technical solutions are more cost- and time-efficient than expert juries (Vukicevic et al. 2022). Consequently, idea evaluation is characterized by a trade-off between efficiency and quality.

Therefore, to leverage the potential of innovation contests, the possibility of replacing expert juries with crowds or LLMs in a way that quality is not impaired has gained interest in prior decision-making literature (Klein and Garcia 2015; Görzen and Kundisch 2016; Magnusson et al. 2016; Mollick and Nanda 2016; Wimbauer et al. 2019; Gao et al. 2023; Peres et al. 2023; Wang et al. 2023).

## 2.2 Crowd-based Idea Evaluation

Crowdsourcing describes open calls for contributions for selected activities to benefit from human collective intelligence and allows for integrating a diverse crowd with different specialized knowledge, backgrounds, and experiences (Howe 2006). When a group of people, i.e. a crowd, evaluates ideas, the value of the work of the individual crowdworkers emerges in the aggregation of all contributions together. Each individual contribution of a crowdworker contributes to the creation of a larger, emergent value that is only realized through the totality and aggregation of all contributions (Geiger et al. 2012). Several studies analyzed replacing expert juries with crowdsourced idea evaluations, for example, in corporate innovation contests (Klein and Garcia 2015; Görzen and Kundisch 2016; Magnusson et al. 2016; Wimbauer et al. 2019) or evaluating projects in the arts industry (Oosterman et al. 2014; Mollick and Nanda 2016). Generally, the results regarding crowd performance are mixed. Oosterman et al. (2014) see significantly better performance in expert juries on an image annotation task. On the contrary, several studies see congruence between expert and crowd evaluations and, thus, see high potential in crowd evaluations—at least for specific domains (Magnusson et al. 2016; Mollick and Nanda 2016; Wimbauer et al. 2019). Groups tend to be more efficient for complex tasks than individuals (Almaatouq et al. 2020). To adequately address crowd evaluations and find generalizable results, an understanding of the underlying evaluation process is necessary.

In sum, existing research established crowd evaluations as a promising way to evaluate ideas (Magnusson et al. 2016). However, research has not yet investigated the potential of crowd evaluations under the consideration of contest specialization. While specialized knowledge has been mentioned as an essential factor for idea evaluation (Görzen and Kundisch 2016), understanding its impact on crowd performance is still needed.

## 2.3 LLM-based Idea Evaluation

LLMs are computational models designed to analyze and generate text (Susarla et al. 2023). LLMs may rely on the Generative Pre-trained Transformer (GPT) architecture, known for its “attention mechanism” (Vaswani et al. 2017). The attention

mechanism enables the model to learn different positions of words in a sentence during the processing of input sequences by assigning individual weights to each element in the input, facilitating strong contextualization (Wolfram 2023). Such models are pre-trained on extensive text datasets, creating a deep representation of the semantic structure of natural language. While LLMs yield fascinating results that are well-documented, the knowledge embedded in such models is broad and not task-specific. However, LLMs can be fine-tuned to specific tasks or domains (Ray 2023) and might be coupled with other knowledge sources such as knowledge graphs.

Several empirical studies have been conducted to investigate the capabilities of LLMs (Kocoń et al. 2023; Zhang et al. 2023). For instance, Kocoń et al. (2023) investigated an LLM's capabilities on 25 different Natural Language Processing (NLP) tasks, such as sentiment analysis, emotion recognition, and stance detection. Their study compares the LLM with the best existing task-specific automation solution. They showed that the LLM, on average, performed 25% worse and concluded that the LLM can only cope with specific tasks to a limited extent. In contrast, Zhang et al. (2023) examine LLMs' capabilities in stance detection. In such tasks, a subject's standpoint (for, against, or neither) to a claim in a text is analyzed. The study found that the LLM can keep up with the performance of task-specific state-of-the-art solutions.

In addition, several studies have been published showing that LLMs have expert level knowledge in domains such as ophthalmology, law or operations management (Terwiesch 2023; Martin et al. 2024; Thirunavukarasu et al. 2024). For example, Martin et al. (2024) benchmark LLMs against a ground truth provided by senior lawyers in the task of contract review. Their findings indicate that LLMs match or surpass human accuracy in identifying legal issues while being significantly faster and more cost-effective. These studies demonstrate that LLMs are already capable of performing many tasks in knowledge work more accurately, efficiently, and cost-effectively than humans. Building on these findings, we aim to explore whether LLMs can also make decisions in more complex and unstructured domains.

Most published studies have been in the zero-shot range, describing an LLM's ability to execute a new task without specific training (Ray 2023). As existing studies demonstrated good LLM performance in zero-shot tasks, there are first indications that LLMs possess zero-shot qualities across domains (Kocmi and Federmann 2023; Kocoń et al. 2023). In addition to zero-shot prompting, more complex prompting strategies, such as few-shot prompting and chain of thought prompting, can increase the quality of LLM output (Kocmi and Federmann 2023).

Even though the examples above do not directly relate to the realm of complex problem idea evaluation, they hint at LLMs' potential in this area. First, they show that LLMs can succeed in zero-shot tasks. Second, the NLP skills examined in these studies include both Natural Language Understanding and Natural Language Generation. These skills are required when evaluating solution ideas to understand the idea itself and to generate an evaluation. Yet, there is also research on LLMs' ability to evaluate text (Gao et al. 2023; Kocmi and Federmann 2023; Wang et al. 2023). Prompting an LLM to rate the quality of text summaries from different datasets showed that an LLM can evaluate text and, in some instances, is closer to human

expert juries than previous technological evaluation models (Gao et al. 2023). While this represents an important step towards leveraging LLMs for evaluation tasks, evaluating solutions for complex problems requires additional specialized domain knowledge (Karvonen and Brand 2009). Investigating the potential of LLMs in this regard is essential because it may help close one of the most severe bottlenecks of innovation contests.

### 3 Hypothesis Development

We investigate whether crowds or LLMs can replace expert juries in idea evaluation for complex problems. Therefore, crowd and LLM performance is assessed as a measure of how close the evaluation is to the expert jury's decisions. An expert jury's decision is not an objective measure of idea quality. However, in the absence of such a measure, the experts' assessment is as close to a ground truth as one can get for the context of our study (Klein and Garcia 2015; Blohm et al. 2016). Our interest lies in analyzing the effect of contest specialization and the evaluation procedure under which each decision maker can be utilized most.

#### 3.1 Effect of Specialization

The type of task decision makers face strongly impacts their evaluation performance (Poole et al. 1985). Contest specialization is a relevant task characteristic because many topics addressed in innovation contests for complex problems are highly specialized and, thus, of little familiarity to a general crowd. For human decision makers, little familiarity with a topic indicates that individuals cannot use long-term memory resources to tackle a task and, thus, are bound to the restraints of the working memory (Kalyuga and Singh 2016). Thus, high contest specialization increases the evaluation task's complexity (Song and Bruning 2016). Consequently, higher complexity leads to worse performance by crowdworkers on a given task (Maynard and Hake 1997). This is caused by high complexity causing a mismatch between available information and task requirements in relation to one's available processing capacity (Cheng et al. 2020). We thus hypothesize:

**H1a** *For absolute rating, crowd evaluation performance is lower for innovation contests of higher specialization.*

As seen in the H1a, our default for examining specialization is the absolute rating of ideas. This approach is also the default for LLM assessments. Generally, LLMs should be able to evaluate due to their ability in zero-shot range (Gao et al. 2023; Kocmi and Federmann 2023; Wang et al. 2023). In addition, previous research also points to the ability to evaluate. Research is less clear regarding the relationship between contest specialization and the evaluation performance of LLMs. LLMs are trained on a large corpus of training data from a broad range of topics and, thus, possess a wide-ranging knowledge base that includes specialized knowledge in many

different domains (Feuerriegel et al. 2023; Ray 2023). With regard to complex tasks, the transformer architecture of LLMs enables them to process complex contexts through different weightings in the input—the attention mechanism (Vaswani et al. 2017). This is critical to an LLM’s ability to answer specialized questions appropriately by considering the nuances of a specialized domain. Consequently, LLMs are probably familiar with highly specialized topics, and their evaluation performance does not depend on specialization levels. We thus hypothesize:

**H1b** *For absolute rating, LLM evaluation performance is not impacted by contest specialization.*

### 3.2 Interaction Effect of Computed Ranking and Specialization

Hypothesizing the negative effect of specialization on crowd performance (H1a), the question of how to counteract this effect emerges. One possibility lies in specifically adapting the evaluation procedure to account for the challenges associated with high contest specialization. Absolute rating performance, as argued in the derivation of H1a, likely decreases with higher specialization. However, even if a decision maker evaluates several ideas independently on an absolute scale, the previous evaluations will subconsciously impact the subsequent evaluation scores (Mussweiler and Englich 2005). In the field of decision-making, this effect is called anchoring or anchoring bias (Tversky and Kahneman 1974), which, in our case, refers to the fact that the first proposals’ evaluations influence the evaluation of subsequent proposals. The resulting subliminal comparison between proposals creates an implicit anchor that helps to decide what constitutes a good or a bad idea. By arranging the absolute evaluation scores from highest to lowest and making decisions based solely on the ranking position rather than the magnitude of the scores, a computed ranking is formed. This computed ranking allows for more consistent evaluations, as the individual evaluations are seen in a context of relative positions. We thus hypothesize:

**H2a** *Computed ranking reduces the negative impact of specialization on crowd performance as compared to absolute rating.*

LLMs are not explicitly developed for evaluation but rather conduct the evaluation based on their broad language-based training data (Ray 2023). These data included in the LLM provide an implicit anchor with which the ideas can be compared. When computing a ranking of the absolute ratings, differences between the individual proposals are presented in a more differentiated way, as they are not only considered in relation to the training data but also in relation to each other. Thus, a computed ranking should increase performance across tasks of every specialization. In line with our assumption in H1b, which posits that the evaluation performance is equal for varying degrees of specialization, we continue to not assume an interaction effect of computed ranking and specialization:

**H2b** *Computed ranking does not impact the effect of specialization on LLM performance.*

### 3.3 Interaction Effect of Relative Ranking and Specialization

In addition to the computed ranking, relative ranking can be used to evaluate ideas. The computed ranking is based on sorting values that were rated on a scale independently. In contrast, relative ranking describes the hierarchical arrangement of ideas (Ovadia 2004). Research has found neither of the two approaches to be superior across use cases (Rankin and Grube 1980). However, for crowdsourced idea evaluations, several studies indicated the superiority of relative ranking evaluations (Görzen and Kundisch 2016; Magnusson et al. 2016; Mollick and Nanda 2016). Magnusson et al. (2016) show that experts and crowds exhibit significantly higher evaluation conformance when using relative ranking. Rating of ideas requires individuals to create their own evaluation schemas and set their own boundaries for good and bad ideas. For relative ranking, the crowd does not need to know what ideas are on either extreme of a scale (Blohm et al. 2016). Instead, it is sufficient for crowdworkers to compare the available ideas regardless of where those ideas are placed in the complete spectrum of good and bad ideas. This suggests that the crowd performance could be improved when applying relative ranking in the presence of specialization. Hence, we hypothesize:

**H2c** *Relative ranking reduces the negative impact of specialization on crowd performance as compared to computed ranking.*

To decide whether an idea is good or bad in absolute terms, an LLM must build an implicit hierarchy based on its knowledge base. To make the task easier for the LLM, users can request it to rank ideas (Ji et al. 2023). This provides the LLM with reference points for its decision. Reference points can be explicitly provided through pairwise comparisons—a form of relative ranking (Gao et al. 2023). Pairwise comparison is particularly suitable for LLMs, as a minimal context window is required. Due to the presence of more than one idea, the model does not have to compare the ideas solely to training data but has two ideas that it can compare. This should make it easier to evaluate the ideas leading to higher performance than the computed ranking. In line with our assumption in H1b, which posits that the evaluation performance is equal for varying degrees of specialization, we continue to not assume an interaction effect of relative ranking and specialization. We thus hypothesize:

**H2d** *Relative ranking does not impact the effect of specialization on LLM performance.*

Figure 1 shows the resulting research model of this study.

Table 1 provides a detailed definition of the key constructs of our research endeavor.

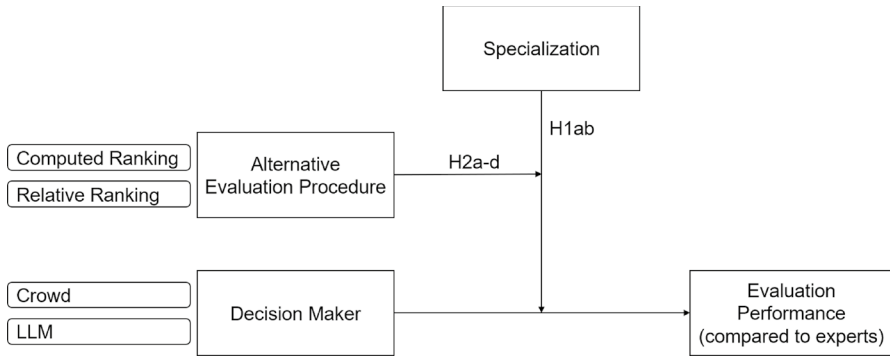


Fig. 1 Research Model

Table 1 Key Constructs of the Paper

Construct	Definition	Source
Decision Maker	The decision maker is the entity that evaluates the idea.	Klein and Garcia (2015)
Specialization	The degree to which a topic belongs to one particular, knowledge-intensive domain rather than a broad field of interest.	(Jeppesen and Lakhani 2010)
Absolute Rating	In absolute rating, the decision maker rates alternatives independently of each other, for example, on a Likert scale. Only those alternatives with a score above a specified threshold are considered for further development; others are rejected.	Ovadia (2004)
Computed Ranking	In computed ranking, the decision maker rates alternatives independently of each other, for example, on a Likert scale. By organizing the absolute evaluation score of the alternatives in ascending order, a ranking is then computed. The decision regarding advancement is now based on the ranking, as opposed to the magnitude of the absolute scores.	Based on Ovadia (2004)
Relative Ranking	In relative ranking, the decision makers accomplish a direct hierarchical arrangement of n alternatives. As a result, a rank is ascribed to each alternative, which serves as the basis for the decision regarding advancement.	Ovadia (2004)
Evaluation Performance	Evaluation Performance is a measure of how close the crowd's and LLM's evaluation is to the expert jury's decision.	Mollick and Nanda (2016)

## 4 Method

To test our hypotheses, we use real-life ideas from innovation contests related to climate change submitted to a platform community called MIT Climate CoLab. On the platform, solution ideas are referred to as *proposals*.

## 4.1 Innovation Contest Context and Expert Evaluation

The MIT Climate CoLab is an open community platform aiming at tackling global climate change via innovation contests through the benefits of collective intelligence. The platform includes more than 125,000 participants who submit proposals addressing diverse facets of climate change. Proposals are mainly submitted by individuals and institutions, such as NGOs, activists, and start-ups. All proposals have the same basic structure and typically include a description of the suggested action, the expected impact, geographic restrictions, and investment needs. In the contests, the semi-finalist and finalist selection are conducted by an expert panel of typically two to five policymakers, business representatives, investors, or scientists. People who generate the proposals and expert jury members who evaluate the proposals do not overlap.

To compare the expert jury with crowd and LLM evaluations, we selected completed contests and proposals hosted not later than 2018, containing at least 30 initial submissions and having detailed expert evaluations. Each contest portrays a sub-problem of the complex problem climate change (Peters 2017). To assess each contest's specialization, we conducted a pre-test with nine researchers and five student assistants interested and knowledgeable in IT-enabled crowdsourcing and climate change. All participants independently ranked six contests according to the perceived specialization of the contests problem definition. They used the contest description and all the proposal titles as a basis for their decision. Intercoder reliability in the form of Kendall's coefficient of concordance (Kendall's W) is 0.79, indicating a strong agreement across raters (Kendall and Smith 1939). The two contests with the highest and the lowest levels of specialization were selected for the subsequent analysis to best portray the extremes of low and high specialization. From lowest to highest specialization, the contests are titled as follows: (1) "Shifting Attitudes and Behaviors", (2) "Adaptation", (3) "Land Use: Agriculture, Forestry, Waste Management", and (4) "Energy Supply." The contests' problems represent complex problems as they involve numerous interconnected causes and dynamic interactions that cannot be solved in a linear way. Changes in attitudes and behavior, adaptation processes, land use and energy supply require consideration of a wide range of social, ecological and economic factors that often conflict with each other and affect different interest groups. For our study, the four contests represent four different treatment groups, each of which differs in the specialization of the problem to be worked on. Since the four groups and the individual proposals are structured in the same way, the four groups differ only in the specialization of their content.

For our study, we selected the proposals and expert evaluations from the semi-finalist selection phase in each of the four contests. We selected this phase because the quality difference between the proposals is higher in this phase (as compared to the final), offering a better first data set for assessing the general potential of the two decision makers. Both proposals that advanced to the semi-final and those that did not were presented to the crowdworkers and the LLM for gathering evaluations. Overall, 26 proposals from each group (104 proposals in total) were presented in the evaluation task to the crowd and the LLM. The original expert jury was asked to evaluate proposals on the four dimensions of novelty, feasibility, impact, and

presentation on a 4-point Likert scale as well as to decide whether they believed the proposal should advance to the next round on a 4-point scale from “absolutely no” to “absolutely yes” (Appendix A). We refer to this overall quality indication as the decision makers evaluation score. We took the same approach for the crowd and the LLM. For each proposal, we assessed for each of the two evaluators if they agree with the expert jury on whether a proposal advances to the next competition round or not.

## 4.2 Crowd Evaluation

To collect crowd evaluations on the selected proposals, we used MTurk as a crowd-sourcing platform. For Information Systems research specifically, Jia et al. (2017) demonstrate the suitability of MTurk populations when conducting empirical studies. We recruited crowdworkers living in the U.S. to ensure English language proficiency (O’Leary et al. 2014). We further implemented validity indicators in the form of response time, response patterns, attention checks, and unusual comments to open questions (O’Leary et al. 2014; Jia et al. 2017). Those who failed attention checks, had unrealistically short response time, or had invalid answers to the open question were marked as fraudulent.

Each crowdworker was presented six proposals from one of the four groups to evaluate the criteria presented above. In a second step, the crowdworkers had to rank the six proposals directly. We did this ranking task for two of the four groups. Proposals were randomly selected from the advancing and not advancing proposals groups and were shown in random order. In our study, we also control for participants’ scientific background and their self-assessment of confidence in their evaluations. Participants received a monetary reward of USD 3.60 for about 25 min of work. Prior research suggests that this moderate level of compensation is adequate on MTurk and encourages valid responses (Jia et al. 2017). In total, our data comes from 200 different participants. 57 participants were identified as fraudulent or bots based on the quality indicators and were thus removed from the dataset. This procedure resulted in 143 remaining participants.

## 4.3 LLM Evaluation

To generate evaluations from the LLM, we used the OpenAI Application Programming Interface (OpenAI 2023). The GPT-3.5-16 k model was used for the study. We used that model for two reasons: Firstly, the context window of 16 k tokens allows us to put in the lengthy proposals as a whole. Secondly, the knowledge base of this model (from September 2021) is close to the crowd data collection time which took place in 2020 which means that the LLM does not have a big knowledge advantage compared to the crowd.

In order to achieve the best possible output, we used various prompt engineering methods (OpenAI 2024). For example, we assigned the persona of an expert

to the LLM and made clear distinctions between different parts of the input (the evaluation task and the proposal). We iteratively developed the prompt until the LLM outputted the numerical evaluation of each individual criterion for the ideas (Kocoń et al. 2023). We used the same prompt for absolute rating and computed ranking. For relative ranking, we included the comparative component between two proposals. Appendix B shows the final prompt for the absolute rating and Appendix C shows the prompt for the relative ranking.

Since we are interested in the basic skills of the GPT model, we left the temperature and all other parameters at their default values (Takagi et al. 2023). Optimization of such parameters could further improve the LLMs' performance. Thus, the results shown here should be considered a lower bound of the LLMs' evaluation potential. Two evaluation rounds were conducted. In the first round, each proposal was individually assessed 30 times based on the four evaluation criteria that the jury used and an overall evaluation score. This resulted in 780 evaluations per group and 3,120 evaluations in total. In the second round, a pairwise comparison was conducted between all possible proposal pairs within each group (Gao et al. 2023; Zheng et al. 2023). Since each group contains 26 proposals, 1,300 pair comparisons were conducted by the GPT model. The cost of evaluating one proposal amounted to \$0.02 for the LLM.

The different methodological choices for measuring relative ranking between crowd and LLM in that regard may be viewed critically. Harmonizing them was discussed but deemed unpractical. There are two main reasons.

- (1) Economic reasons for the crowd: A pairwise comparison by the crowd would result in more effort for the crowd, which costs time and money and thus reduces the economic advantages of using a crowd for platform owners. The limitation that arises from this choice is that we may underestimate the potential of the crowd because they are asked to keep six proposals in mind which is an additional cognitive load.
- (2) Exceeding technical limits of LLM: The evaluation of six proposals by the LLM revealed a position bias in the evaluation. This is a common issue in the literature around LLM as a Judge (e.g., Zheng et al. 2023). Our analysis revealed that the order of the proposals in the prompt correlates strongly with the ranking in the output. This suggests that the LLM weights content that appears earlier in the prompt higher, while later ones are given less consideration.

As a result, we decided that it was the most practical path forward to use two different methods for measuring relative rankings between the crowd and the LLM.

#### 4.4 Evaluation Procedures

To examine specialization and the effect of different evaluation procedures on the effect of specialization, we used the following three evaluation procedures.

*Absolute Rating* For the absolute rating, each decision maker was asked whether the proposal should advance to the final round on a scale from one (absolutely no) to four (absolutely yes). For absolute rating, we split the scale in half to classify proposals that received a mean rating below 2.5 as *not advance* and those 2.5 and above as *advance*.

*Computed Ranking* For the computed ranking, we sorted all proposals by the average evaluation score from best to worst. We classified the best X proposals as *advance*, where X represents the number of proposals that the expert jury advanced to the next round.

*Relative Ranking* For relative ranking, each crowdworker was shown six proposals with the task of ranking them from best to worst. This data was then used to calculate an average ranking for each proposal. The proposals were then sorted according to their average scores. We have implemented the relative ranking for the LLM by pairwise comparison of proposals. We sorted all the proposals based on the number of pair comparison wins from most wins to most minor wins. For both, we then classified the best X proposals as *advance*, where X represents the number of proposals that the expert jury advanced to the next round.

## 4.5 Performance Measurements

Our analysis assesses three performance measurements: balanced accuracy, the positive predictive value (PPV), and the negative predictive value (NPV) (Safari et al. 2015).

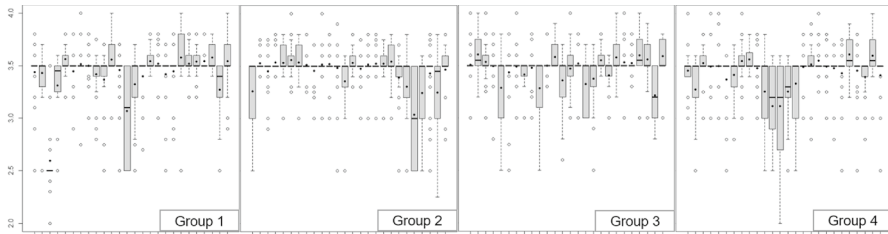
We use *balanced accuracy* as an indicator for general performance of the decision maker, which is defined as the arithmetic mean of true positive rate (proportion of actual positives correctly classified) and true negative rate (proportion of actual negatives correctly classified). Balanced accuracy considers the class distribution and is useful when dealing with imbalanced datasets, which is the case for our sample (Broderson et al. 2010).

The *PPV* reflects the decision makers' reliability when positively rating ideas. It is the number of proposals that the decision maker and the jury classified as *advance* divided by the number of proposals the decision maker classified as *advance* (Safari et al. 2015). The *NPV* assesses how reliable the decision makers' negative evaluations are. It is calculated by dividing the number of proposals that the decision maker and the jury classified as *not advance* divided by the number of proposals the decision maker classified as *not advance* (Safari et al. 2015).

## 5 Results

### 5.1 Demographics and Descriptive Statistics

Of the 143 MTurk participants, 42% are female, 57% male, and 1% chose not to specify. On average, each MTurker spent 25 min on the crowdsourcing task. The GPT model only takes 2 s to evaluate one proposal. For the crowd, we analyze the



**Fig. 2** Boxplots of the Evaluation Scores of the LLM for each of the 104 Proposals

homogeneity of the participants concerning gender, age, education, and profession across the four groups to ensure comparability. For discrete variables, we use  $\chi^2$  tests for homogeneity, and for continuous variables, we use ANOVA (Hair et al. 2010). Based on those tests, participants do not differ significantly between the groups (5% significance level).

Next, we present descriptive statistics for the evaluations of the LLM. Figure 2 illustrates the distribution of the evaluation scores for each proposal across the four groups. The boxplots show the distribution of the 30 evaluation scores for each proposal. The black line represents the median, the black dot represents the mean, and the white dots show outliers.

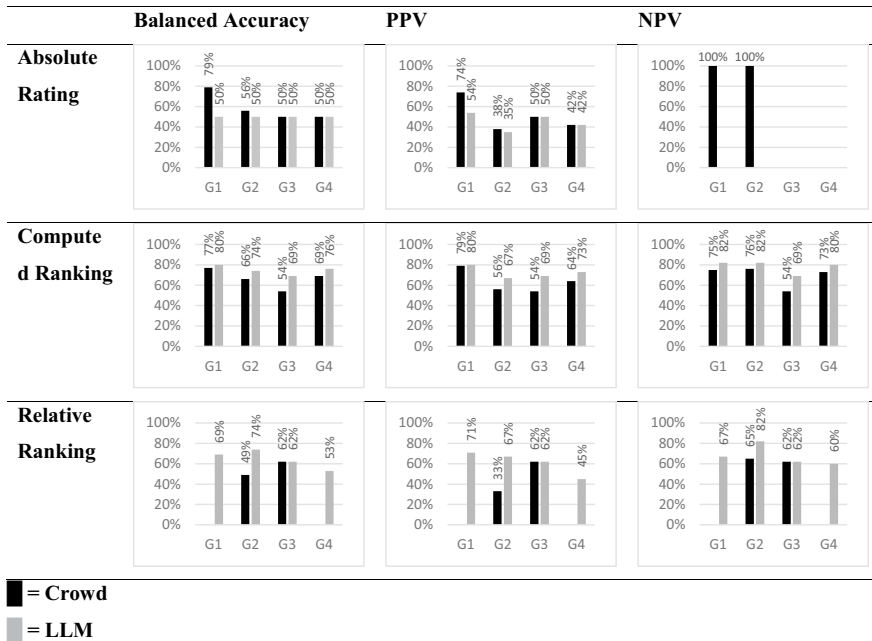
We can observe a strikingly low variance in the evaluation scores in the boxplots, with mean and median values clustered around 3.5. Thus, we conducted an ANOVA to find out whether significant differences exist between the mean evaluation scores of the proposals within each group. We conducted the ANOVA for each treatment group with 26 groups (since each group contains 26 proposals). Group 1 and 2 exhibit significant differences in their evaluation scores ( $p$ -value  $< 0.05$ ). Group 3 and 4 do not show significant differences in their evaluation scores, as their  $p$ -values are  $> 0.05$ . However, the post-hoc Tukey test (Abdi and Williams 2010) shows significant differences between several proposals with regard to the mean evaluation scores for all four groups. This enables an interpretable ranking of proposals, demonstrating that the LLM does not rate all proposals equally.

## 5.2 Hypotheses Tests

Table 2 presents crowd and LLM performance measurements for each group's absolute rating, computed ranking, and relative ranking. Numerically, balanced accuracy is best for each group when using the LLM's computed ranking. Likewise, PPV is best for each group when using the LLM's computed ranking. For NPV, different combinations of evaluation metric and decision maker perform best.

To test the hypotheses, we use two statistical tests. First, we use Fisher's exact test to see if the performance in the groups changes significantly by changing the evaluation procedure. We conduct the test with two evaluation procedures as the rows of the  $2 \times 2$  matrix, and proposals ranked identical as well as proposals ranked differently than the expert jury as its columns (Upton 1992). Second, we use a  $\chi^2$  test for independence to see if there is a significant effect of specialization. The test is

**Table 2** Performance Measurements for Absolute Rating, Computed Ranking and Relative Ranking with ascending Specialization from Group 1 (G1) to Group 4 (G4)



performed in a 4 × 2 matrix with the four groups of one evaluation procedure as the rows. The two columns indicate whether a proposal was evaluated as identical or different from the expert jury as the second variable (McHugh 2013).

**5.2.1 Effect of Specialization on Absolute Rating Performance (H1a and H1b)**

Assessing balanced accuracy values for an absolute rating of the crowd, we see a performance decline from the least specialized group 1 (79%) to the most specialized group 4 (50%). In group 1, the crowd rated ideas rather similar to the expert jury. With rising group specialization, the crowd’s proposal evaluations differ more strongly from the experts. The resulting p-value of a  $\chi^2$  test for independence is 0.015, indicating that the groups’ observed performance is significantly different between at least two of the groups. A pairwise comparison suggests that this is due to significantly higher performance in group 1 compared to the three other groups.

For absolute ratings, *advance* classifications by the crowd were more reliable for the little specialized groups as PPV decreases with group specialization. NPV values are at 100% for groups 1 and 2, indicating that all proposals the crowd rated negatively were also rated the same way by the expert jury. The crowd rated every proposal as advance for the two more specialized groups, which also explains the low PPV in those groups (not applicable (N/A) for NPV). Balanced accuracy and NPV both *support H1a*. The lower scores for PPV as compared to NPV indicate that the crowd a) tends to rate proposals too leniently and marks too many proposals as

*advance*, especially in highly specialized groups. Also, b) the crowd is more reliable in their negative judgments.

Analyzing the balanced accuracy values of the LLM for absolute rating, we see a more consistent picture. Across specialization levels, we can observe a balanced accuracy of 50%. As the performance is constant, we conclude that it is independent of specialization, *supporting H1b*. NPV values for the LLM all show N/A. This results from the LLM marking every single proposal as *advance*. The PPV values are also influenced by the advancement of all proposals. Like the crowd, we observe that too many proposals are evaluated as *advance*. For absolute rating, LLM performance is rather poor, independent from group specialization.

## 5.2.2 Interaction Effect of Computed Ranking and Specialization (H2a and H2b)

The second set of hypotheses concerns the computed ranking process. For the *crowd*, we see a substantial performance increase for balanced accuracy with a delta of up to 19% (group 4) when shifting from absolute rating to computed ranking. Robust across all four groups, PPV values increase between 4% (group 3) and 22% (group 4) compared to absolute rating. In computed ranking, the crowd's *advance* classifications become more reliable. For the less specialized groups, NPV performance decreases as the crowd rates more ideas as *not advance* with a few of them being false classifications. We conduct Fisher's exact test between absolute rating and computed ranking. Performance does not differ significantly between absolute rating and computed ranking ( $p$ -values from group 1 to 4: 1.000, 0.093, 1.000, 0.093).

Next, we assess the interaction effect of computed ranking and specialization. For the crowd, the impact of computed ranking on balanced accuracy is most minor for group 1 (-2%) and largest for group 4 (+19%). Accordingly, with higher specialization, the effect of computed ranking on crowd performance increases ( $\chi^2 p=0.345$ ), indicating that the performance is not significantly different between the groups. Consequently, by changing the evaluation procedure, the specialization effect disappears, which *supports H2a*.

When comparing absolute rating and computed ranking for the *LLM*, a clear increase in performance can be seen across all groups (delta between 19% for group 3 up to 30% for group 1). Fisher's exact test indicates that the distribution of equal and different ratings does differ significantly between absolute rating and computed ranking ( $p=0.075$  for group 1,  $p=0.004$  for group 2,  $p=0.023$  for group 4, with one outlier in group 3:  $p=0.258$ ). This increase is primarily because the GPT model allows all proposals to advance in the absolute rating and is more nuanced in the computed ranking. PPV and NPV also show consistently high values across specialization levels. NPV values are the same or slightly higher than PPV (delta between 0% for group 3 up to 15% for group 2). Thus, the LLM is more reliable in identifying bad proposals. Balanced accuracy for computed ranking is in a rather narrow range between 69% (group 3) and 81% (group 1) and not significantly different ( $\chi^2 p=0.801$ ) across the four groups. Overall, this means computed ranking increases performance across all groups but does not create an effect of specialization on performance, which *supports H2b*.

### 5.2.3 Interaction Effect of Relative Ranking and Specialization (H2c and H2d)

The last set of hypotheses concerns the relative ranking procedure. For the *crowd*, balanced accuracy for the relative ranking is 49% for group 2 and 62% for group 3. Since NPV is equal to or higher than PPV, the crowd is also more reliable in identifying poor proposals than good proposals in the relative ranking. The  $p$ -values of Fisher's exact test between computed ranking and relative ranking are 0.393 for group 2 and 0.779 for group 3, showing that the distribution of equal and different ratings does not differ significantly. The resulting  $p$ -value of a  $\chi^2$  test for independence of 0.779 shows that the observed performance is not significantly different between the groups. Thus, as our results show no further increase in performance, but the specialization effect is also no longer present, *H2c is supported*.

For the *LLM*, balanced accuracy for the relative ranking ranges from 53% (group 4) to 74% (group 2). Analogous to the computed ranking, the NPV values are, on average, better than the PPV values. This means the LLM is also better at identifying poor proposals when applying relative ranking. The difference in balanced accuracy between computed and relative ranking is smallest for group 2 (0%) and highest for group 4 (-23%). The  $p$ -values of Fisher's exact test between computed ranking and relative ranking show that the distribution of equal and different ratings does not differ significantly ( $p$ -values from group 1 to 4: 0.532, 1.000, 0.771, 0.144). To examine the specialization effect, we conduct a  $\chi^2$  test for independence. The resulting  $p$ -value of 0.334 indicates that the groups' observed performance is not significantly different between the groups. Hence, the LLM's rating performance with relative ranking does not depend on specialization, which *supports H2d*.

Table 3 summarizes our hypotheses and the respective empirical results.

**Table 3** Research Hypotheses Overview and Results

#	Description	Empirical Results
H1a	For absolute rating, crowd evaluation performance is lower for innovation contests of higher specialization	Supported
H1b	For absolute rating, LLM evaluation performance is not impacted by contest specialization	Supported
H2a	Computed ranking reduces the negative impact of specialization on crowd performance as compared to absolute rating	Supported
H2b	The computed ranking does not impact the effect of specialization on LLM performance	Supported
H2c	Relative ranking reduces the negative impact of specialization on crowd performance compared to computed ranking	Supported
H2d	The relative ranking does not impact the effect of specialization on LLM performance	Supported

### 5.3 Analysis of Differentiated Proposal Evaluations

To better understand the decision makers' reasons for advancing or not advancing a proposal, we perform a logistic regression of the relationship between the independent variables novelty, feasibility, impact, and presentation (measured on a 4-point Likert scale) and the binary dependent variable (0 for not advance, 1 for advance) (Winship and Mare 1984). As the computed ranking leads to the highest accuracy values, we focus on this evaluation procedure. The assumptions for the logistic regression, particularly the multicollinearity between the independent variables, were checked. We fit 15 regression models in total: For each of the three decision makers, we fit one model per group and an additional overarching model integrating all data. Exponentiated coefficients are used for interpretation, indicating a change in odds (for assessing proposals as advance) when increasing the respective independent variable by one unit. McFadden's Pseudo  $R^2$  is computed as a goodness-of-fit measure, where values around 0.3 represent strong model fit (McFadden 1979). Table 4 presents the results. McFadden's Pseudo  $R^2$  values are high (between 0.347 and 0.694) for the crowd and jury, indicating strong predictability of the decision makers' decision based on the four dimensions. The LLM values are lower (between 0.066 and 0.267), indicating only a moderate fit of the logistic regression model.

For regressions with all groups, all four dimensions significantly and positively impact the decision to classify a proposal as advance. Consequently, for the expert jury, the crowd, and the LLM, a higher evaluation in the individual dimensions leads to a higher probability of advancing the proposal. For the crowd, presentation impacts the decision substantially more than the other dimensions. When looking at the values for all groups of the LLM, presentation, and impact have a high impact,

**Table 4** Logistic Regressions

		All	Group 1	Group 2	Group 3	Group 4
Crowd	Novelty	1.66**	2.02**	3.20**	2.17*	1.05
	Feasibility	1.67**	1.86*	0.68	0.73	4.17**
	Impact	1.69**	2.11**	1.81	0.88	2.56*
	Presentation	4.80**	3.89**	5.30**	8.35**	11.47**
	McFadden's Pseudo $R^2$	0.371	0.368	0.35	0.347	0.598
LLM	Novelty	1.74**	2.07**	1.53**	1.61**	1.47*
	Feasibility	1.57**	1.49	1.05	3.85	3.8**
	Impact	3.48**	3.18**	2.93**	2.88**	11.50**
	Presentation	4.85**	5.35**	2.56**	4.95**	6.42**
	McFadden's Pseudo $R^2$	0.145	0.169	0.066	0.147	0.267
Expert Jury	Novelty	3.54**	5.8*	2.05	5.31*	5.02
	Feasibility	2.59**	10.10**	2.41	1.61	1.12
	Impact	3.11**	9.58**	2.73*	3.90*	1.03**
	Presentation	4.20**	3.65*	3.36*	2.99	3.35*
	McFadden's Pseudo $R^2$	0.515	0.694	0.445	0.468	0.614

\* $p < 0.05$ ; \*\* $p < 0.01$

novelty, and feasibility, a minor impact. In contrast, the expert jury's regression coefficients are more balanced. Analyzing the individual group regression, an effect of specialization can be found for the crowd's evaluation behavior. The more specialized a group is, the more strongly the crowd bases their decision on the proposal presentation rather than impact, novelty, or feasibility. No specialization effects can be identified for the expert jury and the LLM.

## 6 Discussion

### 6.1 Theoretical Implications

Idea evaluation by expert juries is a severe bottleneck in innovation contests, as temporal and financial resources are often scarce. Consequently, we investigate the potential of crowd and LLM evaluations to relieve expert juries of the burden of evaluating solutions to complex problems.

Existing studies on *crowd evaluations* focused on specific idea domains mainly in a corporate setting, and their results regarding crowd performance were mixed (Klein and Garcia 2015; Görzen and Kundisch 2016). Mollick and Nanda (2016) highlight the importance of better understanding the circumstances under which the crowd is a suitable decision maker. Our study provides evidence that contest specialization is a key inhibitor of crowd performance. Accordingly, the crowd performs poorly in the evaluation of highly specialized topics. However, less specialized topics are quite suitable for the crowd, as the crowd is more familiar with these topics and can use their long-term memory resources. Our findings on lenient crowd evaluations and overscoring further indicate that crowdworkers are hesitant to rate a proposal negatively. This effect is particularly strong with high contest specialization, where the presentation of a proposal appears to inappropriately outweigh other relevant characteristics. The crowd likely struggles to reject ideas when they lack the specialized knowledge required to understand what constitutes a good or a poor idea. For research, these contributions shed light on evaluation patterns in crowdsourcing and reveal a potential error source in crowdsourced idea evaluations.

In line with Magnusson et al. (2016), ranking compared to absolute rating can counteract the hesitancy to reject proposals, particularly in the presence of high specialization. Ranking takes the responsibility of rejecting a proposal away from the crowd, as the proposals are sorted in the order of their quality. Contest operators are then responsible for setting a threshold and making the final decision regarding advancing and not advancing ideas. Ranking thus reduces the burden of understanding what good and bad proposals constitute on a universal scale of all existing ideas in favor of a comparative approach between the specific ideas in the respective contest. Therefore, the approach appears promising and highly relevant for innovation contests in the complex problem domain. This highlights that contest operators can affect the crowd's ability to evaluate ideas based on an adequate evaluation procedure design. Our results show that which ranking approach is used is less relevant, as the performance does not differ significantly between the computed ranking and the relative ranking. The original assumption that relative ranking leads to better

performance was not supported here. One possible reason for this could be that the subliminal comparison between proposals in the computed ranking is already sufficient to evaluate the proposals appropriately.

Most previous research on the capabilities of *LLMs* has been related to NLP tasks (Kocoń et al. 2023; Zhang et al. 2023) which we have now extended to a practical evaluation task in the area of complex problems. In line with studies of LLM performance in NLP tasks referring to evaluation (Gao et al. 2023; Kocmi and Federmann 2023; Wang et al. 2023), we show that LLMs possess evaluation capabilities under the right circumstances. The evaluation procedure in particular should be represented by a (relative or computed) ranking, as the LLM performs consistently and accurately when applying a ranking. In contrast, the LLM performs consistently but not necessarily accurately in the absolute rating. A high consistency cannot be used to infer performance because consistency reflects stability, not necessarily the quality of the outcomes. When considering the specialization, the evaluation performance of LLMs seems independent of specialization for each evaluation procedure. This adds to the theoretical foundations of LLMs, showing their inherent capacity to incorporate a wide array of diverse specialized knowledge to carry out evaluation tasks effectively.

A second noteworthy finding is that the LLM in our study tends to assign only very good ratings with minimal variance across proposals on an absolute scale. This suggests that scoring proposals based on an absolute rating may not capture nuanced differences in task complexity and comprehension. At first sight, no clear threshold divides the evaluation scores into good and bad. Comparing the computed with the relative ranking, the LLM performs worse in relative ranking. As this difference has no statistical significance, we cannot say which ranking approach is superior. The computational complexity of the computed ranking is in the class of linear time ( $O(n)$ ), while it is quadratic time ( $O(n^2)$ ) for the relative ranking based on pairwise comparisons. This means that if proposals are compared in pairs, each proposal must be compared with each other and therefore the effort is higher than if you only evaluate a single proposal. For a large number of proposals, this might lead to favoring the computed ranking. Therefore, researchers in the field of LLMs should carefully consider adopting ranking approaches to gain more meaningful insights into the model's performance on different tasks.

Based on a logistic regression, we were able to untangle the basis on which the LLM makes its decision. The moderate fit of the regression models for the LLM shows that we can only explain a small part of the variance of the LLM's decision to advance proposals. LLMs, therefore, remain a kind of black box for us in their mode of operation (Feuerriegel et al. 2023). This finding underlines the importance of Explainable AI to better understand how AI models make decisions (Ray 2023).

Synthesizing the implications of both decision makers, it can be concluded that crowds and LLMs have the ability to evaluate ideas in the complex problem domain. The performance of the LLM is not dependent on the degree of specialization. In contrast, it's a key inhibitor for the crowd, but it can disappear through the choice of evaluation procedures. Further analysis show that both decision makers evaluate the few very best and worst ideas most accurately. Therefore, they are particularly helpful in the preselection phase of innovation contests to weed out bad ideas and reduce the burden on expert juries.

## 6.2 Practical Implications

Several practical implications can be inferred from our study. First, our results show that, in general, crowds and LLMs can be used as decision makers in innovation contests that address complex problems. However, the impact of specialization highlights the need for contest operators to know the specialization of the topic to be evaluated. If information on the specialization is available, a crowd can be used for the evaluation of less specialized areas. If the specialization is not known or critically high, an LLM evaluation might be a more suitable solution as it evaluates regardless of specialization levels.

Second, our findings on the effectiveness of ranking evaluation procedures serve as guidelines for designing evaluation procedures in innovation contests. Innovation contests should use ranking tasks rather than absolute ratings for both crowds and LLMs to achieve assessments closer to expert juries. This holds in particular for evaluations by the crowd in contests where the domain is highly specialized or specialization levels are unknown.

Third, we observe that the presentation of the proposals strongly impacts the decision maker's decision to advance a proposal, particularly for the crowd. This finding could be considered when designing the platform by pointing idea generators to the importance of the presentation and standardizing the presentation's level of detail and professional appearance. Further, the crowd might be explicitly informed not to overly focus on the presentation of a proposal but to focus on aspects such as novelty, feasibility, and impact.

The fundamental challenge of balancing efficiency and quality cannot be solved entirely. However, our study suggests that when an expert jury is not available and the stakes in the decision are not too high, employing LLMs rather than a crowd might be appropriate. We specifically point to LLMs here rather than a crowd as—considering the numerical values of balanced accuracy—the computed ranking from the LLM outperforms other metrics and the crowd (the outperformance is numerical but not statistically significantly different from zero in all cases). The LLM evaluation quality tends not to deteriorate with contest specialization, and—given some technical expertise for assessing an LLM—evaluation is quicker and cheaper with an LLM as compared to a crowd. Further, GPT-3.5 allowed us to achieve roughly 70–80% balanced accuracy compared to the expert jury. Using GPT-4 or other more advanced models that will become available over time will likely increase accuracy. A final word of caution: for high-stakes decisions, 80% accuracy might not suffice, and expert juries might be preferable.

## 6.3 Limitations & Future Research

Our research has limitations, which highlight the need for future research. First, the evaluation of the crowd took place in 2020, and the knowledge base of the LLM extends to 2021. In this period, there might have been technological or political progress in some of the idea areas discussed in the innovation contests. Consequently, these differences may impact the perception of an idea.

Second, our research builds upon the core assumption that expert jury evaluations are closest to the unknown ground truth of idea quality. Accordingly, we assess the decision makers' performance by comparing their evaluation decision to the expert jury. This assumption is based on extensive research in the domain (Klein and Garcia 2015; Blohm et al. 2016; Nagar et al. 2016). However, to fully validate this assumption and better understand the decision makers' evaluation behavior, it would be beneficial for future research to track the implementation progress and resulting benefits of the evaluated ideas. Similar to Wimbauer et al. (2019) research could then assess their actual quality and, even better, evaluate the idea preferences of the decision makers.

After demonstrating the general suitability of LLMs in the complex problem domain, our study offers some pathways for improving evaluation performance. Further research should investigate more complex prompting strategies to achieve higher-quality output. In addition to adapting the prompting, fine-tuning an LLM to the specific task and knowledge area is possible to achieve better performance.

In addition, we are endeavoring to obtain objective evaluations of innovative ideas from an LLM and compare them with expert evaluations. Another relevant research stream deals with the modeling of human opinions by LLMs for specific populations in social science (Argyle et al. 2023; Dominguez-Olmedo et al. 2024; Lee et al. 2024). This research stream is particularly relevant to our work as it contributes to a better understanding of how LLMs can replicate objective and subjective judgments. The research suggests that LLMs could model crowds with diverse opinions that may be well suited to evaluate ideas in combining both LLMs and crowd research. Merging insights from both research streams could open up further perspectives on the evaluation of ideas.

The evolving capabilities of LLMs have been demonstrated in many studies, and it seems inevitable that decision-making will become increasingly automated in the future. This is due to a combination of factors, including technological advances, the growing availability of data and the need for more efficient solutions. As a result, many decision-making processes will increasingly be delegated to machines. Our study highlights that full automation is not yet achievable, underscoring the continued necessity for interaction between humans and technology. Therefore, it is crucial to complement the purely technical perspective with a sociotechnical approach that considers this interplay. Thus, future decision-making literature should take a holistic view of the entire decision-making system, focusing on both effectiveness and efficiency (Storey et al. 2024). It is therefore imperative to address human computer interaction and the effective delegation between the two in the decision-making process in order to optimally utilize their respective strengths (Baird and Maruping 2021).

## 7 Conclusion

Designing idea evaluations in innovation contests for complex problems more efficiently is increasingly important, as traditional expert juries often constitute a severe financial and temporal bottleneck. This study provides the first detailed assessment of the potential of crowdsourced idea evaluations and LLMs' idea evaluations in the complex problem domain. Our results suggest that specialization and evaluation procedures

affect the crowd's potential across contexts and domains. With regard to the LLM, our results show that it can process tasks even in the zero-shot range, regardless of specialization. Concerning practice, we show that to some degree crowds and especially LLMs are indeed suitable for idea evaluations, even in the complex problem domain. Crowds and LLMs are important contributors in tackling the bottleneck of expert juries.

## Appendix A: Survey Items

Novelty (source: Climate CoLab)

Scale	Label
1	Common, mundane, boring
2	Interesting, but not unheard of
3	Unusual, interesting, imaginative
4	Rare, surprising, challenging paradigms

Feasibility (source: Climate CoLab)

Scale	Label
1	Infeasible socially, politically, legally or technically
2	Challenging, feasibility is questionable
3	Acceptable; Objections & barriers partially addressed
4	Appealing; Potential objections & barriers well addressed

Impact (source: Climate CoLab)

Scale	Label
1	Benefits/ impact not clear
2	Limited benefits/small impact
3	Partial solution/ moderate impact
4	Large, direct, & positive impact

Presentation (source: Climate CoLab)

Scale	Label
1	Neither clear, persuasive nor appealing
2	Only 1 of the following 3 applies: Clear, Persuasive, Appealing
3	Only 2 of the following 3 apply: Clear, Persuasive, Appealing
4	All 3 of the following 3 apply: Clear, Persuasive, Appealing

Overall Score (source: e.g. Görzen et al., 2016): Given the quality of the idea proposal you just saw and the rating you scored on all four dimensions, do you think the idea should advance to the next round?

Scale	Label
1	Absolutely no
2	Rather no
3	Rather yes
4	Absolutely yes

## Appendix B: LLM Prompt for Idea Rating

Please provide your evaluation in the standard format: [Five numbers separated by semicolons].

As an expert jury member in the semifinals of an idea contest with the topic "Shifting Attitudes and Behaviour" you evaluate 26 different proposals of which about half should make it to the final.

Please rate the proposal based on the criterion of novelty using a scale from 1 to 4, where 1 represents "Common, mundane, or boring", 2 stands for "Interesting, but not unheard of", 3 indicates "Unusual, interesting, and imaginative", and 4 signifies "Rare, surprising, challenging paradigms". You are allowed to answer in decimal numbers.

Please rate the proposal based on the criterion of feasibility using a scale from 1 to 4, where 1 represents "Infeasible socially, politically, legally or technically", 2 stands for "Challenging, feasibility is questionable", 3 indicates "Acceptable; Objections & barriers partially addressed", and 4 signifies "Appealing; Potential objections & barriers well addressed". You are allowed to answer in decimal numbers.

Please rate the proposal based on the criterion of impact using a scale from 1 to 4, where 1 represents "Benefits/ impact not clear", 2 stands for "Limited benefits/small impact", 3 indicates "Partial solution/ moderate impact", and 4 signifies "Large, direct, & positive impact". You are allowed to answer in decimal numbers.

Please rate the proposal based on the criterion of presentation using a scale from 1 to 4, where 1 represents "Neither clear, persuasive nor appealing", 2 stands for "Only one of the following three applies: Clear, Persuasive, Appealing", 3 indicates "Only two of the following three apply: Clear, Persuasive, Appealing", and 4 signifies "All three of the following three apply: Clear, Persuasive, Appealing". You are allowed to answer in decimal numbers.

Given the quality of the proposal and the rating you scored on all four criteria, do you think the idea should advance to the final round (Advance-Score)? Please answer this question using a scale from 1 to 4 where 1 represents "Absolutely no", 2 indicates "rather no", 3 stands for "rather yes" and 4 signifies "absolutely yes". You are allowed to answer in decimal numbers.

Please provide your rating in the following row, divided by a semicolon: "Novelty of the proposal"; "Feasibility of the proposal"; "Impact of the proposal"; "Presentation of the proposal"; "Advance-Score of the Proposal". So only provide five numbers divided by a semicolon without written text.

In the next section you can find the proposal.

## Appendix C: LLM Prompt for Relative Idea Ranking

Please provide your evaluation in the standard format: [Nine numbers separated by semicolons].

As an expert jury member in the semifinals of an idea contest with the topic "Adaptation" you evaluate 2 different proposals.

Please rate both proposals based on the criterion of novelty using a scale from 1 to 4, where 1 represents "Common, mundane, or boring", 2 stands for "Interesting, but not unheard of", 3 indicates "Unusual, interesting, and imaginative", and 4 signifies "Rare, surprising, challenging paradigms". You are allowed to answer in decimal numbers.

Please rate both proposals based on the criterion of feasibility using a scale from 1 to 4, where 1 represents "Infeasible socially, politically, legally or technically", 2 stands for "Challenging, feasibility is questionable", 3 indicates "Acceptable; Objections & barriers partially addressed", and 4 signifies "Appealing; Potential objections & barriers well addressed". You are allowed to answer in decimal numbers.

Please rate both proposals based on the criterion of impact using a scale from 1 to 4, where 1 represents "Benefits/ impact not clear", 2 stands for "Limited benefits/small impact", 3 indicates "Partial solution/ moderate impact", and 4 signifies "Large, direct, & positive impact". You are allowed to answer in decimal numbers.

Please rate both proposals based on the criterion of presentation using a scale from 1 to 4, where 1 represents "Neither clear, persuasive nor appealing", 2 stands for "Only one of the following three applies: Clear, Persuasive, Appealing", 3 indicates "Only two of the following three apply: Clear, Persuasive, Appealing", and 4 signifies "All three of the following three apply: Clear, Persuasive, Appealing". You are allowed to answer in decimal numbers.

Given the quality of the proposals and the rating you scored on all four criteria, which of the two proposals should advance to the final round?

Please provide your rating in the following row, divided by a semicolon: "Novelty of the first proposal"; "Feasibility of the first proposal"; "Impact of the first proposal"; "Presentation of the first proposal"; "Novelty of the second proposal"; "Feasibility of the second proposal"; "Impact of the second proposal"; "Presentation of the second proposal"; "seven figure number of proposal that advances". So only provide nine numbers divided by a semicolon without written text.

In the next two sections you can find the two proposals. Each section starts with the seven-figure number of the proposal:

**Acknowledgements** We acknowledge funding by the Federal Ministry of Education and Research, the Bavarian State Ministry of Science and Art, the Ministry of Science, Research and Arts of Baden-Württemberg, and the Hessian Ministry of Higher Education, Research, Science and the Arts for the ABBA project (Grant Numbers 16DHBKI002, 16DHBKI003, 16DHBKI005). We submitted an extended abstract of this paper solely focused on crowdsourcing to the 6th International Conference on Computational Social Science hosted by the MIT in Boston. After the acceptance of our contribution, we presented our research idea and preliminary results at the conference in July 2020. We gained valuable feedback (e.g., regarding our theoretical framework), which we integrated into our paper. Further, we extended the manuscript by a LLM perspective to account for recent technological developments and enrich the contribution.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This work was supported by Bundesministerium für Bildung und Forschung (Grant Nos. 16DHBKI002, 16DHBKI003, 16DHBKI005), Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg (Grant No. 16DHBKI002), Hessisches Ministerium für Wissenschaft und Kunst, (Grant No. 16DHBKI003), and the Bavarian State Ministry of Sciences and Art (Grant No. 16DHBKI005).

## Declarations

**Conflict of interest** The authors have no relevant financial or non-financial interest to disclose.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Abdi H, Williams LJ (2010) Newman–Keuls test and Tukey test. Encyclopedia of research design. SAGE Publication. <https://doi.org/10.4135/9781412961288.n266>
- Almaatouq A, Yin M, Watts DJ (2020) Collective problem-solving of groups across tasks of varying complexity. Preprint
- Argyle LP, Busby EC, Fulda N, Gubler JR, Rytting C, Wingate D (2023) Out of one, many: using language models to simulate human samples. *Polit Anal* 31:337–351. <https://doi.org/10.1017/pan.2023.2>
- Baird A, Maruping LM (2021) The next generation of research on IS use: a theoretical framework of delegation to and from agentic IS artifacts. *MISQ* 45:315–341. <https://doi.org/10.25300/MISQ/2021/15882>
- Blohm I, Leimeister JM, Krcmar H (2013) Crowdsourcing: how to benefit from (too) many great ideas. *MISQ Exec* 4:199–211
- Blohm I, Riedl C, Füller J, Leimeister JM (2016) Rate or trade? Identifying winning ideas in open idea sourcing. *Inf Syst Res* 27:27–48. <https://doi.org/10.1287/isre.2015.0605>
- Brodersen KH, Ong CS, Stephan KE, Buhmann JM (2010) The balanced accuracy and its posterior distribution. In: Brodersen KH, Ong CS, Stephan KE, Buhmann JM (eds) The balanced accuracy and its posterior distribution. *IEEE*, 3121–3124. <https://doi.org/10.1109/ICPR.2010.764>
- Cheng X, Fu S, de Vreede T, de Vreede G-J, Seeber I, Maier R, Weber B (2020) Idea convergence quality in open innovation crowdsourcing: a cognitive load perspective. *J Manage Inf Sys* 37:349–376. <https://doi.org/10.1080/07421222.2020.1759344>
- Chesbrough H (2006) Open innovation: a new paradigm for understanding industrial innovation. In: Chesbrough H, Vanhaverbeke W, West J (eds) Open innovation: researching a new paradigm. Oxford University Press, Oxford, pp 1–14
- Dominguez-Olmedo R, Hardt M, Mandler-Dünner C (2024) Questioning the survey responses of large language models. In: Proceedings of the 38th conference on neural information processing systems
- Feuerriegel S, Hartmann J, Janiesch C, Zschech P (2023) Generative AI. *Bus Inf Syst Eng* 66:111–126. <https://doi.org/10.1007/s12599-023-00834-7>
- Fischer A, Greiff S, Funke J (2012) The Process of solving complex problems. *J Prob Solv* 4:19–42. <https://doi.org/10.7771/1932-6246.1118>
- Gao M, Ruan J, Sun R, Yin X, Yang S, Wan X (2023) Human-like Summarization Evaluation with ChatGPT. Preprint. <https://doi.org/10.48550/arXiv.2304.02554>
- Geiger D, Rosemann M, Fieft E, Schader M (2012) Crowdsourcing information systems—definition, typology, and design. proceedings of the 33th international conference on information systems
- Gerlach J, Buxmann P, Diney T (2019) “They’re all the same!” Stereotypical thinking and systematic errors in users’ privacy-related judgments about online services. *J. Assoc. Inf. Syst.* 20:787–823. <https://doi.org/10.17705/1jais.00551>

- Gimpel H, Graf-Drasch V, Laubacher RJ, Wöhl M (2020) Facilitating like darwin: supporting cross-fertilisation in crowdsourcing. *Decis Support Syst* 132:113282. <https://doi.org/10.1016/j.dss.2020.113282>
- Görzen T, Kundisch D (2016) Can the crowd substitute experts in evaluation of creative ideas? An experimental study using business models. In: *Proceedings of the 22nd Americas Conference on Information Systems*
- Hair JF, Black WC, Babin BJ, Anderson RE (2010) *Multivariate data analysis: a global perspective*, 7th edn. Pearson
- Han Y, Ozturk P, Nickerson J (2020) Leveraging the wisdom of the crowd to address societal challenges: Revisiting the knowledge reuse for innovation process through analytics. *J Assoc Inf Syst.* 21:1128–1152. <https://doi.org/10.17705/1jais.00632>
- Head B (2008) Wicked problems in public policy. *Public Policy* 3:101–118
- Howe J (2006) The rise of crowdsourcing. *Wired Mag* 14:1–4
- Jeppesen LB, Lakhani KR (2010) Marginality and problem-solving effectiveness in broadcast search. *Organ Sci* 21:1016–1033. <https://doi.org/10.1287/orsc.1090.0491>
- Ji Y, Gong Y, Peng Y, Ni C, Sun P, Pan D, Ma B, Li X (2023) Exploring ChatGPT's ability to rank content: a preliminary study on consistency with human preferences. Preprint. <https://doi.org/10.48550/arXiv.2303.07610>
- Jia R, Steelman Z, Reich BH (2017) Using Mechanical Turk data in IS research: Risks, rewards, and recommendations. *Comm Ass Inform Syst.* 41:301–318. <https://doi.org/10.17705/1CAIS.04114>
- Kalyuga S, Singh A-M (2016) Rethinking the boundaries of cognitive load theory in complex learning. *Educ Psychol Rev* 28:831–852. <https://doi.org/10.1007/s10648-015-9352-0>
- Karvonen A, Brand R (2009) Technical expertise, sustainability, and the politics of knowledge. In: Kitting G, Lipschutz RD (eds) *Environmental governance: power and knowledge in a local-global world*. Routledge, London, pp 38–59
- Kendall MG, Smith BB (1939) The problem of m rankings. *Ann Math Stat* 10:275–287
- Klein M, Garcia ACB (2015) High-speed idea filtering with the bag of lemons. *Decis Support Syst* 78:39–50. <https://doi.org/10.1016/j.dss.2015.06.005>
- Kocmi T, Federmann C (2023) Large language models are state-of-the-art evaluators of translation quality. Preprint. <https://doi.org/10.48550/arXiv.2302.14520>
- Kocoi J, Cichecki I, Kaszyca O, Kochanek M, Szydło D, Baran J, Bielaniec J, Gruza M, Janz A, Kanclerz K, Kocoi A, Koptyra B, Mieszczonko-Kowszewicz W, Miłkowski P, Oleksy M, Piasecki M, Radliński L, Wojtasik K, Woźniak S, Kazienko P (2023) ChatGPT: Jack of all trades, master of none. *Inform Fusion* 99:101861. <https://doi.org/10.1016/j.inffus.2023.101861>
- Lee S, Peng T-Q, Goldberg MH, Rosenthal SA, Kotcher JE, Maibach EW, Leiserowitz A (2024) Can large language models estimate public opinion about global warming? An empirical assessment of algorithmic fidelity and bias. *PLOS Clim.* <https://doi.org/10.1371/journal.pclm.0000429>
- Magnusson PR, Wästlund E, Netz J (2016) Exploring users' appropriateness as a proxy for experts when screening new product/service ideas. *J Prod Innovat Manag* 33:4–18. <https://doi.org/10.1111/jpim.12251>
- Martin L, Whitehouse N, Yiu S, Catterson L, Perera R (2024) Better call GPT, comparing large language models against lawyers. Preprint. <https://doi.org/10.48550/arXiv.2401.16212>
- Maynard DC, Hakel MD (1997) Effects of objective and subjective task complexity on performance. *Hum Perform* 10:303–330. [https://doi.org/10.1207/s15327043hup1004\\_1](https://doi.org/10.1207/s15327043hup1004_1)
- McFadden D (1979) Quantitative methods for analyzing travel behavior of individuals: some recent developments. In: Hensher DA, Stopher PR (eds) *Behavioural travel modelling*. Croom Helm, London, pp 279–318
- McHugh ML (2013) The chi-square test of independence. *Biochem. med.* 23:143–149. <https://doi.org/10.11613/bm.2013.018>
- Mollick E, Nanda R (2016) Wisdom or madness? Comparing crowds with expert evaluation in funding the arts. *Manage Sci* 62:1533–1553. <https://doi.org/10.1287/mnsc.2015.2207>
- Mussweiler T, English B (2005) Subliminal anchoring: Judgmental consequences and underlying mechanisms. *Organ Behav Hum Dec* 98:133–143. <https://doi.org/10.1016/j.obhdp.2004.12.002>
- Nagar Y, De Boer P, Garcia ACB (2016) Accelerating the review of complex intellectual artifacts in crowdsourced innovation challenges. In: *Proceedings of the 37th international conference on information systems.* <https://doi.org/10.5167/uzh-126367>

- O'Leary M, Wilson J, Metiu A (2014) Beyond being there: the symbolic role of communication and identification in perceptions of proximity to geographically dispersed colleagues. *MISQ* 38:1219–1244. <https://doi.org/10.25300/MISQ/2014/38.4.13>
- Oosterman J, Bozzon A, Houben G-J, Nottamkandath A, Dijkshoorn C, Aroyo L, Leyssen MH, Traub MC (2014) Crowd vs. experts: Nichesourcing for knowledge intensive tasks in cultural heritage. In: Proceedings of the 23<sup>rd</sup> international conference on world wide web, 567–568. <https://doi.org/10.1145/2567948.2576960>
- OpenAI (2023) API-Reference. <https://platform.openai.com/docs/api-reference>. Accessed 12 November 2023
- OpenAI (2024) Prompt Engineering. <https://platform.openai.com/docs/guides/prompt-engineering/strategy-write-clear-instructions>. Accessed 21 June 2024
- Ovadia S (2004) Ratings and rankings: reconsidering the structure of values and their measurement. *Int J Soc Res Method* 7:403–414. <https://doi.org/10.1080/1364557032000081654>
- Özaygen A, Balagué C (2018) Idea evaluation in innovation contest platforms: a network perspective. *Decis Support Syst* 112:15–22. <https://doi.org/10.1016/j.dss.2018.06.001>
- Peres R, Schreier M, Schweidel D, Sorescu A (2023) On ChatGPT and beyond: how generative artificial intelligence may affect research, teaching, and practice. *Int J Res Mark* 40:269–275. <https://doi.org/10.1016/j.ijresmar.2023.03.001>
- Peters BG (2017) What is so wicked about wicked problems? A conceptual analysis and a research program. *Policy Soc* 36:385–396. <https://doi.org/10.1080/14494035.2017.1361633>
- Poole MS, Seibold DR, McPhee RD (1985) Group decision-making as a structural process. *Q J Speech* 71:74–102. <https://doi.org/10.1080/00335638509383719>
- Rankin WL, Grube JW (1980) A comparison of ranking and rating procedures for value system measurement. *Eur J Soc Psychol* 10:233–246. <https://doi.org/10.1002/ejsp.2420100303>
- Ray PP (2023) ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *I O t and Cyber-Phys Syst* 3:121–154. <https://doi.org/10.1016/j.iotcps.2023.04.003>
- Rittel HW, Webber MM (1973) Planning problems are wicked. *Pol Sci* 4:155–169
- Safari S, Baratloo A, Elfil M, Negida A (2015) Evidence based emergency medicine part 2: positive and negative predictive values of diagnostic tests. *Emerg J* 3:87–88
- Simon HA (1960) *The New Science of Management Decision*. Harper, New York
- Song M, Bruning R (2016) Exploring effects of background context familiarity and signaling on comprehension, recall, and cognitive load. *Educ Psychol-UK* 36:691–718. <https://doi.org/10.1080/01443410.2015.1072133>
- Stern NH (2006) *The economics of climate change: The Stern review*, 1st edn. Cambridge Univ. Press, Cambridge
- Storey VC, Hevner AR, Yoon VY (2024) The design of human-artificial intelligence systems in decision sciences: a look back and directions forward. *Decis Support Syst* 182:114230. <https://doi.org/10.1016/j.dss.2024.114230>
- Susarla A, Gopal R, Thatcher JB, Sarker S (2023) The Janus effect of generative AI: charting the path for responsible conduct of scholarly activities in information systems. *Inform Syst Res* 34:399–408. <https://doi.org/10.1287/isre.2023.ed.v34.n2>
- Takagi S, Watari T, Erabi A, Sakaguchi K (2023) Performance of GPT-3.5 and GPT-4 on the Japanese Medical Licensing Examination: Comparison Study. *JMIR Med Educ* 9. <https://doi.org/10.2196/48002>
- Terwiesch C (2023) *Would Chat GPT Get a Wharton MBA?* Mack Institute for Innovation Management at the Wharton School, University of Pennsylvania
- Thirunavukarasu AJ, Mahmood S, Malem A, et al. (2024) Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: a head-to-head cross-sectional study. *PLOS Digit. Health*. <https://doi.org/10.1371/journal.pdig.0000341>
- Tversky A, Kahneman D (1974) Judgment under uncertainty: heuristics and biases. *Science* 185:1124–1131. <https://doi.org/10.1126/science.185.4157.1124>
- Upton GJG (1992) Fisher's exact test. *J R Stat Soc A Stat* 155:395–402. <https://doi.org/10.2307/2982890>
- Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A, Kaiser Ł, Polosukhin I (2017) Attention is All you Need. In: Proceedings of the 31<sup>st</sup> international conference on neural information processing systems
- von Hippel E (1994) “Sticky information” and the locus of problem solving: Implications for innovation. *Manage Sci* 40:429–439. <https://doi.org/10.1287/mnsc.40.4.429>

- Vukicevic A, Vukicevic M, Radovanovic S, Delibasic B (2022) BargCrEx: a system for bargaining based aggregation of crowd and expert opinions in crowdsourcing. *Group Decis Negot* 31:789–818. <https://doi.org/10.1007/s10726-022-09783-0>
- Wang J, Liang Y, Meng F, Sun Z, Shi H, Li Z, Xu J, Qu J, Zhou J (2023) Is ChatGPT a Good NLG Evaluator? A preliminary study. Preprint. <https://doi.org/10.48550/arXiv.2303.04048>
- Wimbauer L, Figge P, Häussler C (2019) Distant search, but local implementation? Using the crowd's evaluation to overcome organizational limitations in the selection of crowdsourced ideas. In: Proceedings of the 40<sup>th</sup> international conference on information systems
- Winship C, Mare RD (1984) Regression models with ordinal variables. *Am Sociol Rev* 49:512–525. <https://doi.org/10.2307/2095465>
- Wolfram S (2023) What Is ChatGPT Doing ... and Why Does It Work? <https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-ng-and-why-does-it-work/>. Accessed 19 November 2023
- Zhang B, Ding D, Jing L (2023) How would Stance Detection Techniques Evolve after the Launch of ChatGPT? Preprint. <https://doi.org/10.48550/arXiv.2212.14548>
- Zheng L, Chiang W-L, Sheng Y, Zhuang S, Wu Z, Zhuang Y, Lin Z, Li Z, Li D, Xing EP, Zhang H, Gonzalez JE, Stoica I (2023) Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In: Proceedings of the 37<sup>th</sup> conference on neural information processing systems

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## Authors and Affiliations

**Henner Gimpel**<sup>1,2,3</sup> · **Robert Laubacher**<sup>5</sup> · **Fabian Probst**<sup>1,2,3</sup> · **Ricarda Schäfer**<sup>1,4</sup> · **Manfred Schoch**<sup>1,3,6</sup>

✉ Fabian Probst  
fabian.probst@fim-rc.de

Henner Gimpel  
henner.gimpel@uni-hohenheim.de

Robert Laubacher  
rjl@mit.edu

Ricarda Schäfer  
ricarda.schaefer@fim-rc.de

Manfred Schoch  
manfred.schoch@fit.fraunhofer.de

<sup>1</sup> FIM Research Center for Information Management, Augsburg, Germany

<sup>2</sup> University of Hohenheim, Stuttgart, Germany

<sup>3</sup> Branch Business & Information Systems Engineering of the Fraunhofer FIT, Augsburg, Germany

<sup>4</sup> University of Augsburg, Augsburg, Germany

<sup>5</sup> Massachusetts Institute of Technology, Cambridge, MA, USA

<sup>6</sup> Esslingen University of Applied Sciences, Esslingen, Germany