Evaluating AI-Generated Solution Ideas: A Comparative Study of AI and Human Assessments for Sustainable Process Design

Mas'udah[1], Pavel Livotov[1], Saptadi Nugroho[1,2]

[1] Offenburg University of Applied Sciences, Badstr. 24, 77652 Offenburg, Germany
[2] Albert Ludwig University of Freiburg, Georges-Köhler-Allee 101, 79110 Freiburg im Breisgau, Germany
masudah@hs-offenburg.de
pavel.livotov@hs-offenburg.de
saptadi.nugroho@hs-offenburg.de

Abstract. The integration of generative artificial intelligence (AI) in sustainable process design has gained substantial traction, with AI increasingly employed to generate innovative solutions. However, the efficacy of these AI-generated ideas requires rigorous evaluation to ensure their quality. This study examines the dual role of GPT-4o in generating and evaluating solution ideas for sustainable process design at the concept development stage. Focusing on process engineering, the research applies these methods to a case study involving froth flotation for nickel recovery. By comparing AI-driven assessments with those from human experts, the research aims to determine the alignment between AI and human evaluations across key criteria: novelty, feasibility, usefulness, and sustainability. The results reveal strong alignment in most areas, though notable discrepancies in novelty suggest that human expertise remains essential for nuanced judgments on uniqueness. These findings highlight GPT-4o's potential as a preliminary evaluation tool, while also underscoring the need for a hybrid approach that combines AI insights with human expertise.

Keywords: Generative AI; Sustainability; Process Design; AI-Human Evaluation.

1 Introduction

Recent advancements in artificial intelligence (AI) have had a transformative impact across various sectors, enabling rapid innovation and the generation of solutions for complex challenges. Among these developments, Generative Pre-trained Transformer (GPT) models developed by OpenAI [1], have demonstrated significant potential in automating ideation processes, which is especially valuable in the context of sustainable process design. The versatility of these models lies in their ability to process extensive datasets and generate contextually relevant responses, making them well-suited for early-stage concept development aimed at addressing environmental and technical issues [2, 3].

Despite the promise of generative AI, challenges remain in ensuring the quality and applicability of AI-generated solutions. Current approaches often rely heavily on human evaluation to assess key criteria, such as novelty, feasibility, usefulness, and sustainability. These human-driven evaluations, while effective, are time-consuming and often subjective, which highlights the need for AI systems capable of autonomously assessing the quality of their outputs.

Prior studies on generative AI have explored its application in eco-innovation, but limitations persist regarding the AI's ability to evaluate its outputs. This research addresses this gap by investigating GPT-4o's capacity not only to generate but also to evaluate solution ideas. The

research aimed to explore how effectively GPT-4o could autonomously assess its outputs and compare these AI-driven evaluations with those conducted by human experts. By concentrating on established evaluation criteria such as novelty, feasibility, usefulness, and sustainability, the study sought to determine the extent to which GPT-4o's assessments aligned with expert evaluations. The research seeks to address the following questions:

1) Can GPT-4o autonomously evaluate its generated ideas?
2) How do AI evaluations align with human expert assessments across key metrics in sustainable process design?

2 Background and Related Work

2.1 Generative AI in Sustainable Innovation

Generative AI, such as GPT models, has emerged as a powerful tool in the drive toward sustainable innovation by enabling the rapid generation of ideas that address pressing environmental and technical challenges [4]. Since the introduction of GPT-2 in 2019, these models have evolved significantly, with GPT-3 and subsequent versions, including GPT-4o, demonstrating advanced capabilities for generating complex and contextually relevant ideas from extensive datasets [1, 5, 6]. This versatility makes GPT models particularly valuable in fields where rapid innovation and sustainability are essential, such as process engineering, where early-stage concept development plays a pivotal role in tackling environmental issues and technical obstacles.

Despite the promise of generative AI, significant challenges remain in ensuring the quality and applicability of AI-generated solutions. Assessing these qualities has largely depended on human evaluations, which present scalability issues and introduce subjectivity into the assessment process. For instance, Zu et al. [2, 3] utilised GPT-3 to generate biologically inspired design concepts but relied solely on a design team to evaluate feasibility and novelty. While human expertise is invaluable, it often requires substantial time and resources, thus limiting the potential for large-scale, efficient evaluations. Similarly, Li et al. [7] explored multiple generative AI tools, including ChatGPT (GPT-3.5) [1], Midjourney (via Discord) [8], and Stable Diffusion (via WebUI) [9], to facilitate sustainable design projects. Although the study incorporated topics aligned with Sustainable Development Goals (SDGs), it focused exclusively on usability and depended merely on human assessment. While these previous studies successfully demonstrate that generative AI can support sustainable innovation, they fall short in exploring AI-driven evaluation mechanisms, leaving a significant gap in objectively assessing the quality of AI-generated ideas.

Building on this foundation, the previous research by authors [4], integrated nature-inspired principles into a GPT-3.5 model to generate eco-innovative solutions for challenges in process design. This study not only used GPT-3.5 to generate solution ideas but also to evaluate them through self-assessment, with comparisons drawn against human evaluations. However, the study was limited by a lack of in-depth sustainability assessments. The evaluation metrics were relatively general and did not delve into the specifics, leaving a gap in comprehensively assessing the AI-generated solution ideas. This gap underscores the need for more robust evaluation frameworks that incorporate detailed sustainability metrics, enabling AI to perform more rigorous self-assessments in alignment with sustainable development goals.

2.2 Idea Evaluation

Evaluating ideas is an essential component of the innovation process, as it helps in selecting concepts that are most likely to succeed and contribute meaningfully to their intended goals. Common criteria for idea evaluation typically include novelty, feasibility, usefulness, and sustainability [10–12]. These criteria provide a structured basis for assessing the quality of ideas and determining their potential impact.

Various methods have been developed to evaluate these criteria, ranging from expert assessments to crowdsourcing and AI-assisted frameworks. Dean et al. [11] focused on measuring ideas through constructs such as novelty, workability, relevance, and specificity using multi-dimensional scales rated by human experts. The study noted that human raters often interpret these criteria differently, leading to variability in results. This variability underscores a broader challenge in maintaining consistency across evaluations, especially when relying on subjective human assessments.

In another study, Baffo et al. [12] employed utility value analysis to score ideas based on environmental, economic, and social sustainability parameters, allowing for a comprehensive sustainability assessment. This method provides a structured approach to sustainability but tends to be labour-intensive and primarily suited for detailed project-level evaluations, making it less practical for rapid, large-scale screening in early-stage concept development.

Furthermore, a recent study by Mesbah et al. [13] integrated crowdsourced evaluations with machine learning models to improve scalability and consistency. In this framework, crowd raters are guided by well-defined scales to measure criteria such as novelty and usefulness. The machine learning component adjusts for variability in crowd ratings, providing a more reliable aggregate score. However, while the study enhances scalability, it remains limited by the variability and potential biases introduced by crowd raters.

3 Methodology

3.1 Research Design

This study employed a dual approach in which GPT-4o was used to both generate and evaluate solution ideas for sustainable process design, focusing on a specific case study outlined in Table 1. GPT-4o was selected for its enhanced contextual understanding and advanced generative capabilities [14]. Compared to previous versions, GPT-4o has demonstrated improved performance in generating contextually relevant responses to complex prompts, making it particularly suitable for generating innovative solutions [1, 14].

Table 1. Case study: problem and ideal final result

Case study	Problem description	Ideal final result
Froth Flotation for Nickel Recovery	The utilisation of chemicals in the process results in water pollution and the generation of solid waste. Additionally, the low efficiency of the process contributes to elevated production expenses.	Develop a sustainable and effective procedure that reduces the utilisation of chemicals and waste, thereby decreasing production expenses.

3.2 Solution Idea Generation

In alignment with the methodology outlined in the authors' prior research [4], this study employed a solution-driven approach (SDA) in biomimetics, combined with AI-automated prompting to generate solution ideas. The SDA framework leverages principles derived from natural ecosystems exposed to environmental stress, such as mangroves, rainforests, caves, etc. as a basis for developing innovative solutions. By analysing the adaptation strategies within these ecosystems, the approach identifies inventive principles that can be applied to engineering challenges. AI-automated prompting complemented the basic prompts, guiding GPT-4o to achieve more refined results by structuring the input queries and enhancing the output iteratively. The idea generation process followed four key phases, as outlined in Table 2.

Table 2. Idea generation using SDA with generative AI [4]

SDA Phase	Basic prompting	AI-automated prompting
1. Identification and analysis of the natural solution in ecosystem exposed to environmental stress (e.g. mangroves, rainforest, etc.)	Identify the key components and adaptation strategies that help the following ecosystems survive in hostile environments, and explain how they aid survival under environmental stress. *[list all-natural ecosystems that are intended to be identified and analysed]*	Basic prompting for phase 1 + Follow the instructions below: 1) Revise the prompt to be clear, concise and easily understood by you. Ask any relevant questions needed to improve the prompt. 2) Execute the revised prompt upon approval, and provide iterative feedback to refine the information generated. Ask: 'Do you agree with the revised prompt (please type NO or YES)?' If 'NO', ask for more detail to refine the prompt. If 'YES', proceed to provide the answer, and then ask: 'Do you want more [e.g. information, ideas, examples, etc] (please type MORE) or go to the next step (type NEXT)?' If 'MORE', suggest 5 more [e.g. information, ideas, examples, etc] and repeat the feedback question. If 'NEXT', ask: 'What else can I assist you?' Wait for the feedback. If I give you an inquiry, you repeat instructions 1 to 2 above until the job is done.
2. Extraction of natural solution principles	Based on the adaptation strategies identified, distil abstract, non-biomimetic principles that capture the essence of how these ecosystems survive in hostile environments. Name these principles succinctly, ensuring they are generalized enough to apply across different engineering challenges.	
3. Ideas generation with nature-inspired principles to solve the problem	Using the extracted inventive principles, generate 5 ideas that address the sustainable process design challenge to the following problem. Ensure that the ideas are distinct and not direct copies of natural processes. *[Describe the problem and the desired result]*	
4. Concepts generation	Create 5 inventive solution concepts by combining different complementary ideas to address the problem. These solution concepts should offer comprehensive solution approaches that incorporate various elements to enhance useful action and mitigate harmful effects.	

In Phase 1, GPT-4o was prompted to identify and analyse adaptation strategies within selected ecosystems. The initial basic prompts were designed to elicit information about the ecosystem components and their survival mechanisms. The AI-automated prompting then provided additional structured steps, enabling GPT-4o to iteratively refine the prompts based on feedback, ensuring a deeper and more accurate understanding of each ecosystem's adaptation strategies.

In Phase 2, the focus shifted to extracting natural solution principles from these adaptation strategies. GPT-4o was directed to distil these principles into abstract, non-biomimetic forms that could be applied across various engineering contexts. The AI was prompted to avoid direct copying of natural components and instead translate the underlying inventive principles into a generalised format that could inspire innovative solutions.

Phase 3 involved using these nature-inspired principles to generate specific solution ideas addressing the sustainable process design challenge outlined in the case study. GPT-4o was directed to use these principles to produce a diverse range of ideas, generating up to 50 ideas to address the problem. These ideas were then evaluated using the criteria outlined in Section 3.3.

In the final phase, GPT-4o was tasked with creating inventive solution concepts by combining complementary top-scoring ideas from phase 3. These concepts incorporated various elements to enhance beneficial actions and mitigate harmful effects, aligning closely with sustainability goals. In this study, GPT-4o was asked to create up to five solution concepts to address the problem. To ensure consistency and minimise variations in AI responses, both generation and evaluation of the idea were conducted on the same day and at the same time.

3.3 AI-generated Solution Idea Evaluation

The AI-generated solution ideas were evaluated by both AI and human experts using metrics as shown in Table 3. The definitions and rationale for these criteria were derived from the literature [3, 10, 11]. A total of 50 solution ideas for the case study were assessed using these metrics on a scale of 0 to 2.

- *AI self-evaluation* – ChatGPT was asked to evaluate its solution ideas.
- *Human-assisted evaluation* – Two experts, specialising in process engineering and database system engineering, were trained to perform a manual evaluation of the solution ideas independently.

Table 3. Assessment criteria for generated solution idea

Parameter	Description	Rating scale
Novelty - assesses the uniqueness of the solution idea from existing solutions		
• Originality	The solution introduces a unique approach that is not found in existing solutions, such as those found in patent databases, published articles, and other sources	0 - Not novel (common or existing solution) 1 - Moderately novel (introduces some new aspects) 2 - Highly novel (completely new or unique)
• Inventiveness	The solution combines existing concepts in new ways	
• Paradigm shift	The solution transforms existing norms, conventions, or models	

Feasibility - evaluates the viability of implementing the solution idea		
• Technical viability	The necessary technologies and resources are available for implementation	0 - Unviable (highly impractical) 1 - Moderately feasible (possible but requires effort) 2 - Highly feasible (easily implementable)
• Financial viability	The solution has reasonable cost implications compared to the expected benefits	
• Scalability	The solution can be effectively adapted for large-scale or varying needs	
Usefulness - measures the solution idea's potential to address the identified problem effectively		
• Effectiveness	The idea directly addresses the core issues	0 - Useless (does not address the problem) 1 - Moderately useful (resolves a few issues) 2 - Highly useful (completely addresses the problem)
• Practicality	The solution is applicable in real-world settings or targeted process	
• Relevance	The solution meets the expectations or needs of users	
Sustainability - estimates the environmental, social and economic impact of the solution idea		
• Environmental impact	The solution minimises harm to the environment	0 - Unsustainable (significant negative impacts) 1 - Moderately sustainable (minor negative impacts) 2 - Highly sustainable (major positive impacts)
• Social impact	The solution contributes positively to social well-being	
• Economic impact	The solution provides economic benefits and promotes financial stability	

To assess the level of agreement between AI and human evaluators, Cohen's Kappa [15] was used to measure inter-rater agreement between individual human evaluators and GPT-4o. This statistic provides insight into how consistently the AI ratings align with each human rater independently. Additionally, Fleiss' Kappa [16] was applied to evaluate overall agreement across all evaluators, including GPT-4o and the two human experts, thereby offering a comprehensive measure of consensus. These analyses help quantify the reliability of the AI's evaluations in comparison to human judgments across the assessment categories.

Table 4. Interpretation of kappa values [17]

Kappa value	Agreement reliability
≤0	No agreement
0.1 – 0.20	Slight Agreement
0.21 – 0.40	Fair Agreement
0.41 – 0.60	Moderate Agreement
0.61 – 0.80	Substantial Agreement
0.81 – 1.00	Almost Perfect Agreement

4 Results and Discussion

4.1 Idea evaluation

Figure 1 compares AI-driven and human expert evaluations across four criteria: novelty, feasibility, usefulness, and sustainability. Overall, the results show a strong alignment between AI and human assessments, particularly in feasibility, usefulness, and sustainability, supporting GPT-4o as a preliminary evaluation tool where human judgment is resource-intensive. A notable difference appears in novelty, where AI-rated ideas are higher than human evaluators. This suggests that GPT-4o perceives greater originality, likely due to its ability to draw upon a vast dataset and combine existing concepts in new ways. In contrast, human experts may apply stricter criteria, expecting more distinct deviations from conventional solutions. This divergence underscores a need for refinement to bring AI's novelty assessments closer to human standards. These results imply that, while AI can be a valuable asset for initial assessments, human input remains essential for a more nuanced understanding, particularly in the evaluation of novelty.

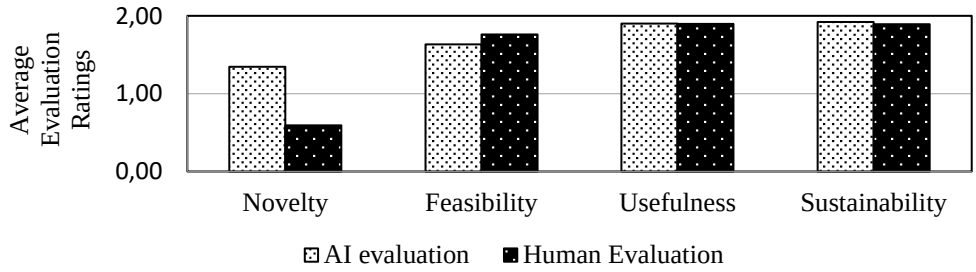

Fig. 1. The average evaluation rating of generated ideas: AI vs human

To further illustrate these findings, Table 5 presents a sample of ideas generated by GPT-4o to address the challenges in the case study, along with their practical implications. The table compares AI and human expert evaluations across key assessment categories. Overall, the results reveal a high level of alignment between AI and human evaluations, with closely matching ratings in most areas, except for novelty, where some discrepancies were observed. This outcome underscores GPT-4o's potential as an effective preliminary evaluation tool. While AI closely mirrors human assessments in feasibility, usefulness, and sustainability, the differences in novelty indicate that human expertise remains essential for nuanced judgment. These findings highlight the value of a hybrid approach that combines AI-driven assessments with human insights, particularly for evaluating aspects such as originality, inventiveness, and paradigm shift.

Table 6 presents the inter-rater agreement across key assessment categories using Cohen's Kappa and Fleiss' Kappa. Overall, the strongest alignment is observed in sustainability, particularly in the environmental and social aspects, where Cohen's Kappa values show higher consistency. In contrast, lower agreement scores are found in novelty, especially for originality and inventiveness, indicating differences in how AI and humans assess uniqueness. Fleiss' Kappa values confirm this trend, showing that sustainability has the highest overall agreement, while novelty displays more variability. These results suggest that GPT-4o aligns well with human assessments in endurance-related aspects but may require refinement in evaluating novelty.

Table 5. Generated idea examples and practical implications for case study according to AI and Human ratings on novelty (N), feasibility (F), usefulness (U), and sustainability (S) (in fragment)

ID	Idea description	Practical implication	Rated by	N				F				U				S				Total
				Originality	Inventiveness	Paradigm shift	Average ($\bar{N}$)	Technical	Financial	Scalability	Average ($\bar{F}$)	Effectiveness	Practicality	Relevance	Average ($\bar{U}$)	Environmental	Social	Economic	Average ($\bar{S}$)	
1	Use biodegradable frothing agents that break down naturally after use	Reduces chemical pollution in discharge water	AI	2	2	2	**2**	2	2	2	**2**	2	2	2	**2**	2	2	2	**2**	**8**
			Human rater 1	1	1	1	**1**	2	1	1	**1.33**	1	2	2	**1.67**	2	2	2	**2**	**6**
			Human rater 2	0	0	0	**0**	1	1	1	**1**	1	2	2	**1.67**	2	2	2	**2**	**4.67**
…	…	…	…	..	..	..		..	..	..		..	..	..		..	..	..	..	..
50	Implement a multi-stage flotation process	Reduce chemical dosages and enhance ore recovery rates	AI	1	1	1	**1**	2	2	2	**2**	2	2	2	**2**	2	2	2	**2**	**7**
			Human rater 1	0	0	0	**0**	2	2	2	**2**	2	2	2	**2**	2	2	2	**2**	**6**
			Human rater 2	0	0	0	**0**	2	1	2	**1.67**	2	2	2	**2**	2	2	2	**2**	**5.67**

Table 6. Inter-rater agreement among raters on novelty (N), feasibility (F), usefulness (U), and sustainability (S) using Cohen's kappa and Fleiss' kappa

Ratings	Cohen's kappa value											
	N			F			U			S		
	Originality	Inventiveness	Paradigm shift	Technical	Financial	Scalability	Effectiveness	Practicality	Relevance	Environmental	Social	Economic
AI – Human rater 1	0.160	0.022	0.239	0.393	0.517	0.506	0.407	0.638	-0.056	0.396	0.558	0.694
AI – Human rater 2	0.093	-0.057	0.153	0.132	0.322	0.330	0.225	0.260	-0.056	0.396	0.457	0.390
Human rater 1- Human rater 2	0.680	0.701	0.696	0.651	0.786	0.737	0.675	0.485	0.728	0.811	0.779	0.336
Fleiss' kappa value (Overall)	**0.191**	**0.164**	**0.212**	**0.362**	**0.541**	**0.471**	**0.442**	**0.464**	**0.250**	**0.556**	**0.606**	**0.481**

4.2 Limitation and Future Work

This study demonstrates GPT-4o's effectiveness as a preliminary evaluation tool, though some limitations are evident. The AI showed discrepancies in assessing novelty, possibly due to its limited ability to fully interpret human standards for uniqueness. Additionally, the reliance on a single AI model and a specific case study restricts the generalisability of the findings, as results may vary with different models or contexts. The limited sample of two human experts for comparison also suggests that a larger, more diverse panel would provide a more comprehensive benchmark for AI assessments. Furthermore, the solution-driven approach (SDA) used in this study employed AI-automated prompting, where ChatGPT revised each prompt based on user instructions. While effective, this iterative process requires manual confirmation at each step. In the future, developing a system like a large language model (LLM) prompter could automate this refinement, reducing manual intervention and improving efficiency.

Future research could explore other AI models and apply the evaluation framework to multiple case studies to broaden the scope of findings. Integrating AI tools trained on creative or sustainability-specific datasets might improve AI's accuracy in evaluating subjective criteria like originality. Furthermore, a hybrid evaluation model that combines AI with a larger, diverse panel of human experts could enhance understanding of AI-human alignment, particularly in complex areas such as originality and paradigm shifts. Expanding stakeholder involvement to include sustainability practitioners and domain experts could also enhance the applicability of AI-based evaluations in sustainable process design.

5 Conclusion

This study finds that GPT-4o can autonomously evaluate its generated ideas, showing strong alignment with human assessments in feasibility, usefulness, and sustainability. However, discrepancies in novelty assessment indicate that human expertise is essential for nuanced judgments on originality, inventiveness, and paradigm shift. These results demonstrate that GPT-4o can function as a preliminary evaluation tool, but combining AI-driven assessments with human insights provides a more robust framework. This study suggests that while GPT-4o aligns well with human evaluations in most criteria, further refinement is needed for novelty assessments. Future research should explore broader applications of AI in sustainable innovation and focus on refining AI models to improve alignment in subjective evaluations like originality.

References

1. ChatGPT Homepage, https://chat.openai.com/, last accessed 2024/09/24.
2. Zhu, Q., Zhang, X., Luo, J.: Generative Pre-Trained Transformers for Biologically Inspired Design. In: 34th International Conference on Design Theory and Methodology (DTM), American Society of Mechanical Engineers (2022). https://doi.org/10.1115/DETC2022-90366. 3.
3. Zhu, Q., Zhang, X., Luo, J.: Biologically inspired design concept generation using generative pre-trained transformers. Journal of Mechanical Design 145(4), art. 041409 (2023).
4. Mas'udah, Livotov, P.: Nature's lessons, AI's power: sustainable process design with generative AI. In: Proceedings of the Design Society, vol. 4, pp. 2129–2138, Cambridge University Press (2024). DOI: https://doi.org/10.1017/pds.2024.215

5. Radford, A., et al.: Language Models are Unsupervised Multitask Learners. Technical report, OpenAi (2019).
6. Brown, T. et al..: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, 33, pp. 1877-1901 (2020).
7. Li, M. et al.: Generative AI for Sustainable Design: A Case Study in Design Education Practices. In: Kurosu, M., Hashizume, A. (eds) Human-Computer Interaction. HCII 2024. Lecture Notes in Computer Science, vol 14687. Springer, Cham (2024). https://doi.org/10.1007/978-3-031-60441-6_5
8. Midjourney Homepage: https://discord.com/, last accessed 2024/09/24.
9. Stable Diffusion Homepage: https://stablediffusionweb.com/WebUI, last accessed 2024/09/24.
10. Ma, K., Grandi, D., McComb, C., Goucher-Lambert, K.: Conceptual Design Generation Using Large Language Models. In: Proceedings of the ASME Design Engineering Technical Conference, vol 6, Boston, Massachusetts, USA (2023). https://doi.org/10.1115/detc2023-116838.
11. Dean, D.L., Hender, J.M., Rodgers, T.L., Santanen, E.L.: Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. Journal of the Association for Information System, vol. 7, Iss.10 (2006). https://doi.org/10.17705/1jais.00106.
12. Baffo, I., Leonardi, M., Bossone, B., Camarda, M.E., D'Alberti, V., Travaglioni, M.: A decision support system for measuring and evaluating solutions for sustainable development. Sustainable Futures, vol. 5, 100109 (2023). https://doi.org/10.1016/j.sftr.2023.100109.
13. Mesbah, S., Arous, I., Yang, J., Bozzon, A., Bozzon, A.: HybridEval: A Human-AI Collaborative Approach for Evaluating Design Ideas at Scale. In: Proceedings of the ACM Web Conference 2023. Association for Computing Machinery, pp. 3837–3848, New York, NY, United States (2023). https://doi.org/10.1145/3543507.3583496
14. Shahriar, S. et al.: Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. (2024). https://doi.org/10.48550/arXiv.2407.09519
15. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20 (1), 37–46. https://doi.org/10.1177/001316446002000104
16. Fleiss, J.L.: Measuring nominal scale agreement among many raters. Psychological Bulletin, 76, 378-382 (1971).
17. Landis, J.R. and Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics, 33(1), 159-174 (1977). http://dx.doi.org/10.2307/2529310.