Proceedings of the Design Society, Volume 6: DESIGN 2026
https://doi.org/10.1017/pds.2026.10603

Multi-agent generative AI for concept evaluation: consistency, knowledge integration and human alignment

Mas'udah[1,✉], Pavel Livotov[1], Björn R. Kokoschko[2,3], Wanyu Xu[4], Immanuel Hendra[5] and Niklas Hartmann[1]

[1] Offenburg University of Applied Sciences, Germany, [2] Otto von Guericke University Magdeburg, Germany, [3] Anhalt University of Applied Sciences, Germany, [4] Texas A&M University, United States of America, [5] Singapore University of Technology and Design, Singapore

✉ masudah@hs-offenburg.de

ABSTRACT: Early-stage concept evaluation is critical for selecting viable designs. This study introduces a multi-agent generative AI framework for assessing concepts across four configurations: AI with retrieval-augmented knowledge, AI without external knowledge, human experts, and a hybrid approach. The findings show that AI panels tend to produce uniform evaluation patterns, while retrieval-augmented knowledge alters rating behaviour without leading to closer alignment with human judgement. Hybrid setting achieved closest alignment, indicating AI is effective when combined with expert interpretation.

KEYWORDS: multi-agent system, artificial intelligence (AI), retrieval-augmented knowledge, concept evaluation, sustainable design

1. Introduction

The early phases of chemical and process engineering design are essential for determining the technical and practical viability of emerging solution concepts. During this stage, multiple designs are typically generated and assessed to identify those with the greatest potential for implementation. Traditional evaluation approaches predominantly rely on expert judgement, often involving panels composed of individuals from various engineering domains. While such human-based assessments provide valuable practical insight, they can be time-consuming, resource-intensive, and their outcomes may vary due to subjectivity, personal bias, disciplinary perspectives, and limited access to knowledge beyond the expertise of individual evaluators (Blessing & Chakrabarti, 2009; Edwards et al., 2025; Ullman, 2003). These challenges become more evident when decisions must address multiple evaluation criteria such as usefulness, feasibility, and sustainability.

Recent developments in artificial intelligence (AI), particularly the emergence of generative language models such as Generative Pre-trained Transformers (GPT) developed by OpenAI (2025), have introduced new opportunities to support early design activities. Generative AI has recently been explored as a complement to expert evaluation, offering structured reasoning and the ability to integrate external knowledge. However, questions remain about the reliability, stability and human alignment of AI-based assessments. In response to this challenge, this study introduces a multi-agent generative AI framework that functions as a virtual expert panel for evaluating early-stage solution concepts. Although subjectivity and cognitive biases are inherent to human-centred assessments, this study does not aim to eliminate or quantify such effects directly. Instead, it focuses on evaluation consistency, the influence of retrieval-augmented knowledge and alignment with human expert judgement. To guide the investigation, the study formulates the following research questions:

https://doi.org/10.1017/pds.2026.10603 Published online by Cambridge University Press

- RQ1: How consistent are the evaluations produced by the multi-agent configuration across domain-specific evaluators?
- RQ2: How does retrieval-augmented knowledge influence the evaluation behaviour and overall panel outcomes of the AI-based configuration?
- RQ3: How closely do the AI-only and hybrid configurations align with human expert judgement in early-stage concept evaluation?

2. Related work

2.1. Concept evaluation in early-phase process design

Evaluating solution concepts in the early stages of process design plays a critical role in determining which options are technically feasible, practically implementable, and aligned with defined design objectives (Blessing & Chakrabarti, 2009; Mesbah et al., 2023). The evaluation is typically carried out to identify the most promising concepts for further development, often under conditions of limited data, time constraints, and conflicting design requirements. To overcome these challenges, several frameworks have been proposed to support systematic evaluation in conceptual design. Common approaches include scoring matrices, decision tables, and weighted evaluation models (Chen et al., 2022; Dym et al., 2013; Pahl et al., 2007), which enable evaluators to assess solution concepts based on predefined criteria such as novelty, feasibility, usefulness, and sustainability (Baffo et al., 2023; Dean et al., 2006). Although such methods improve structure and transparency, they still depend heavily on expert judgement. As a result, evaluations may vary due to subjective judgement, inconsistent application of criteria and disciplinary differences among evaluators (Blessing & Chakrabarti, 2009; Edwards et al., 2025; Ullman, 2003). These persistent challenges highlight the need for approaches that reduce inconsistency while still respecting the inherent subjectivity of design evaluations. Rather than eliminating human subjective insight, digital and AI-assisted methods aim to support more stable, knowledge-informed reasoning during early-stage assessment.

2.2. Generative AI in design concept assessment

Recent advances in generative AI have created new opportunities to support early-stage concept assessment. Several studies have begun examining the ability of large language models (LLMs) to approximate or complement human evaluative judgement. Xu et al. (2024) and Mas'udah et al. (2024) measured inter-rater agreement between ChatGPT and expert panels and reported moderate consistency but notable variability in contextual reasoning. Edwards et al. (2025) examined the use of vision-language models (VLMs) as AI judges in design and found that the best-performing model achieved expert-level agreement for certain metrics, such as uniqueness and drawing quality, but did not consistently reach human-expert equivalence across all evaluation criteria. Moreover, hybrid approaches that combine AI-generated insights with expert validation (Mas'udah et al., 2026, 2025; Mesbah et al., 2023; Sauer & Burggräf, 2025) have shown potential for improving assessment quality but still face challenges related to consistency, transparency and robustness. Existing applications typically rely on single-agent models, which limits the diversity of perspectives and contextual grounding available during evaluation. In this context, recent developments in multi-agent generative AI and retrieval-augmented reasoning offer promising opportunities to strengthen early-stage evaluation.

2.3. Multi-agent AI and RAG for evaluation

Multi-agent generative AI has emerged as a promising approach for improving reasoning quality by distributing tasks among specialised agents. Frameworks such as AutoGen (Wu et al., 2023) and multi-agent judge architectures (Guo et al., 2024; Qian et al., 2025; Rasheed et al., 2024) demonstrate improved robustness and reduced bias compared with single-model assessments. However, despite these advances, multi-agent systems have not yet been applied to structured concept evaluation in engineering design, where assessments must account for criterion-specific and domain-specific requirements.

In parallel, retrieval-augmented generation (RAG), an approach in which an AI model retrieves relevant external domain knowledge and integrates it into its reasoning process, further enhances the factual grounding and contextual relevance of generated outputs (Lewis et al., 2021). In engineering design, previous studies have applied RAG to enhance problem formulation, idea generation, and solution optimisation (Mas'udah et al., 2026, 2025). Although these studies demonstrate that external domain

https://doi.org/10.1017/pds.2026.10603 Published online by Cambridge University Press

knowledge can improve the quality and relevance of generative outputs, RAG has not been used for structured concept evaluation, where evidence-based justification and domain alignment are essential. Current research therefore lacks a framework that integrates multi-agent reasoning with retrieval-augmented knowledge for early-stage engineering concept evaluation. Addressing this gap requires empirical analysis of evaluation consistency, the influence of retrieved evidence and the extent to which AI-based assessments align with human expert judgement.

3. Methodology

3.1. Research design

This study employs a multi-agent generative evaluation framework developed using GPT-5.1, the most recent OpenAI model available at the time of experimentation (November 2025). Figure 1 illustrates the overall workflow of the proposed framework which consists of two main stages.

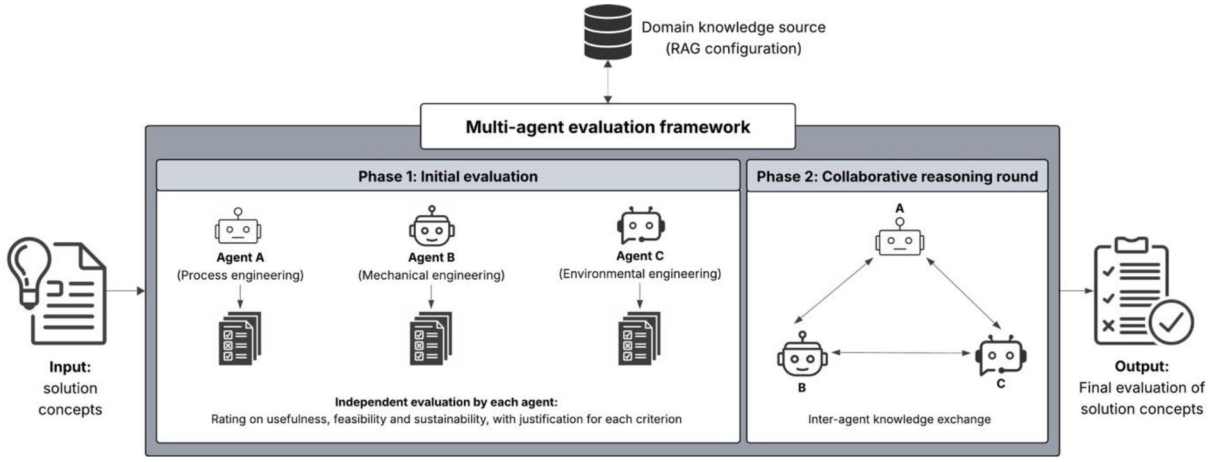


Figure 1. Multi-agent generative AI framework for solution concept evaluation

The first stage of the framework focuses on the independent evaluation of solution concepts generated for the case study (Section 3.2). Three role-specialised agents are instantiated from the GPT-5.1 model, representing process engineering (Agent A), mechanical engineering (Agent B), and environmental engineering (Agent C). Each agent receives the same concept description and independently evaluates the concept based on three criteria: usefulness, technical feasibility, and environmental sustainability. The evaluation criteria follow established early-stage design assessment frameworks and are aligned with the assessment structure applied in earlier work on AI-assisted concept optimisation (Mas'udah et al., 2026, 2025), where usefulness and feasibility are considered fundamental dimensions of early-stage concept assessment (Blessing & Chakrabarti, 2009; Dean et al., 2006; Ullman, 2003), while sustainability has become essential in modern process and environmental engineering evaluation frameworks (Klüppel, 2005; United Nations, 2015). Collectively, these criteria provide a balanced representation of functional benefit, implementability, and environmental responsibility. A standardised five-point scale (1, 2, 3, 4, 5) was applied to evaluate each concept's usefulness, technical feasibility and environmental sustainability, with 1 representing the lowest and 5 the highest score. This rating structure follows the assessment scheme used in (Livotov et al., 2026; Livotov & Mas'udah, 2025). Their ordinal structure enables evaluators to express the degree to which a concept is perceived as useful, feasible, or sustainable, providing a systematic and transparent basis for early-stage design decision-making. Although the scoring framework is identical across agents, each agent applies domain-specific reasoning determined by its assigned role. In the retrieval-augmented condition, the agents have access to an external domain knowledge source during evaluation, as summarised in Table 1. To ensure transparency and reproducibility, the RAG knowledge base consists exclusively of openly accessible engineering documents that are directly relevant to the membrane distillation (MD) case study presented in Section 3.2. These resources were selected to align with the domain responsibilities of the three agents. In the non-RAG condition, the agents rely solely on the knowledge embedded within the GPT-5.1 model. In the second stage of the framework, the agents participate in a collaborative reasoning round. The individual evaluations produced in the first stage are exchanged among the agents, enabling them to review, compare, and refine their reasoning. This inter-agent knowledge exchange supports the

https://doi.org/10.1017/pds.2026.10603 Published online by Cambridge University Press

identification of agreement, divergence, and cross-criterion conflict, resulting in a consolidated final evaluation that reflects integrated multi-perspective reasoning rather than a single-agent judgement. To enable systematic comparison, four experimental settings were examined:

a) **Multi-agent AI with RAG**
The multi-agent system evaluates all concepts with access to external domain knowledge.
b) **Multi-agent AI without RAG**
The same multi-agent system operates using only its internal model knowledge.
c) **Human expert panel**
Three human experts, corresponding to the same disciplinary roles as the AI agents (process, mechanical, and environmental engineering), independently assess the concepts and subsequently discuss their evaluations to produce a consolidated panel outcome. This alignment ensures that comparisons between AI and human evaluations are conducted within an equivalent disciplinary structure.
d) **Hybrid AI-human evaluation**
Human experts review and refine the outputs produced by the AI with RAG setting, which provides the most complete and evidence-supported AI reasoning for human interpretation.

To assess the reproducibility and stability of the model outputs, each AI-based evaluation was executed three times under identical prompting and configuration settings, enabling analysis of run-to-run variation. This research design provides a structured foundation for analysing the influence of role specialisation, retrieval-augmented knowledge, collaborative reasoning, and human expertise on early-stage concept evaluation. The four experimental settings allow systematic comparison of evaluation accuracy, consistency, and reasoning quality across AI-based, hybrid, and human expert assessments.

Table 1. Summary of retrieval-augmented knowledge sources per agent role

Agent role	Domain focus	Knowledge type (RAG)	Representative open-access sources
Process Engineering	MD mechanisms, heat/mass transfer, desalination processes	Principles of MD operation, performance factors, process limitations	- Principles and applications of MD (Drioli, 2017) - MD for water desalination and purification (Camacho et al., 2013) - Desalination technology (Salinas-Rodríguez & Schippers, 2021)
Mechanical Engineering	Module design, materials, mechanical constraints	Material properties, thermal-mechanical stability, module configuration	- NIST materials standard (Elliott et al., 2025) - MD module configuration (Francis et al., 2022) - Membrane patents (Dahiru Umar et al., 2022; Wail Sulaiman & Lawal, 2022)
Environmental Engineering	Environmental impacts, energy use, sustainability indicators	Sustainability frameworks, LCA principles, resource use	- United Nations 2030 Agenda (United Nations, 2015) - LCA framework (ISO 14040, 2006) - LCSA guidelines (UNEP, 2011)

3.2. Case study

To validate the proposed framework, Liquid Gap Membrane Distillation (LGMD) is employed as the reference system for this case study due to its strong coupling between thermal gradients, membrane characteristics, and geometric layout, features that make it suitable as a testbed for early-stage optimisation assessment. LGMD is a thermally driven desalination process in which a heated saline feed flows along a hydrophobic membrane, while a cooler stream behind a condensation surface maintains the temperature gradient. A thin liquid-filled gap between the membrane and condensation plate facilitates vapour condensation and supports compact operation. This configuration, typically operating at 50 - 100°C, offers high salt rejection and compatibility with low-grade heat sources, as demonstrated in (Mas'udah et al., 2026).

Despite these advantages, LGMD performance is constrained by several interconnected issues. First, conductive heat transfer across the liquid gap lowers thermal efficiency and reduces the effective temperature gradient. Second, temperature polarisation at the membrane interfaces further constrains

https://doi.org/10.1017/pds.2026.10603 Published online by Cambridge University Press

vapour transport by diminishing the driving force. Third, the system is prone to membrane wetting and fouling, particularly when processing high-salinity or organic-rich feed streams. These challenges collectively affect permeate flux, energy consumption, and long-term operational stability.
To address these limitations, three solution concepts as presented in Table 2 have been developed in (Mas'udah et al., 2026). These concepts target thermal inefficiencies, flow maldistribution, and membrane performance constraints through modifications in gap design, coolant circulation, and membrane structuring. These solutions serve as the basis for evaluation under the four experimental settings described in Section 3.1.

Table 2. Solution concepts for improving LGMD performance

ID	Solution concept name	Description
SC1	Dual-layer thermal-selective liquid gap	Introduces a dual-layer gap structure in which a low-conductivity fluid (e.g., silicone-based coolant) forms the primary thermal barrier, while a thin high-conductivity sublayer adjacent to the condensation surface enhances condensation. This configuration aims to reduce conductive heat losses while maintaining efficient vapour-phase heat transfer.
SC2	Counter-flow microstructure coolant channel plate	Applies patterned microchannels in a counter-flow arrangement to enhance heat removal and achieve a more uniform temperature distribution. The microstructures promote uniform temperature distribution, minimise thermal dead zones, and stabilise the driving force for vapour transport.
SC3	Hierarchical hydrophobic-superhydrophobic membrane surface	Modifies the membrane by integrating multi-scale surface texturing (micro-pillars combined with nano-roughness) to suppress wetting and reduce fouling adhesion. The modified surface improves vapour-liquid interface stability and maintains performance under high-salinity or organic-rich feeds.

3.3. Evaluation

To assess the performance of the proposed multi-agent evaluation framework, the analysis focuses on measurable indicators aligned with the three research questions.

- **Consistency of evaluation (RQ1)**
  To examine the internal consistency of the evaluations produced by the three domain-specific agents, inter-rater agreement was calculated for each configuration. Weighted Cohen's kappa was used to assess pairwise agreement between the evaluators, as it accounts for ordinal ratings on a Likert scale (Cohen, 1960). Fleiss' kappa (Fleiss, 1971) was applied to examine overall agreement across the three evaluators.

Table 3. Interpretation of kappa values (Landis & Koch, 1977)

Kappa value	Agreement reliability
≤0	No agreement
0.01 – 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

- **Effect of retrieval-augmented knowledge (RQ2)**
  To examine the effect of retrieval-augmented knowledge, the AI without RAG and AI with RAG configurations were compared in terms of their panel consensus scores and the variability of these scores across solution concepts and experimental runs. Panel consensus scores refer to the final ratings agreed upon after the collaborative reasoning stage (Phase 2). Differences in mean scores

https://doi.org/10.1017/pds.2026.10603 Published online by Cambridge University Press

indicate whether RAG influences overall evaluation patterns, while changes in standard deviation reflect its effect on evaluation stability.

- **Comparative performance of AI, human, and hybrid panels (RQ3)**
  To examine how closely AI-based assessments align with human expert judgement, two quantitative indicators were employed. Spearman rank correlation (Spearman, 1904) was applied to assess the similarity in ranking patterns across the three solution concepts, while the mean absolute error (MAE) was used to quantify numerical deviations between each configuration and the human panel (Chai & Draxler, 2014). Both metrics were calculated using the panel consensus scores for each configuration.

4. Results and discussion

4.1. Consistency of evaluation (RQ1)

Table 4 presents the inter-rater agreement results for usefulness evaluations across the four experimental configurations. The AI with RAG configuration exhibits negative pairwise and overall kappa values, reflecting low statistically meaningful consistency among the evaluators. The AI without RAG and Hybrid configurations show near-constant ratings, leading to undefined pairwise agreement coefficients due to insufficient rating variability, while the overall kappa values remain negative. In contrast, the Human configuration demonstrates slight agreement, suggesting greater diversity in expert judgement and more differentiated but less consistent usefulness assessments. Overall, the findings indicate that AI-based evaluators showed highly uniform usefulness ratings, leading to apparent uniformity but limited measurable inter-rater agreement.

Table 4. Inter-rater agreement for usefulness evaluations

Configuration	Avg. Cohen's κ	Fleiss' κ	Interpretation
AI with RAG	−0.058	−0.125	No agreement in both pairwise and overall measures
AI without RAG	–	−0.038	Near-constant ratings with no overall agreement
Human expert	0.121	0.110	Slight agreement in both pairwise and overall measures
Hybrid	–	−0.038	Near-constant ratings with no overall agreement

Turning to feasibility evaluation, Table 5 shows the inter-rater agreement results across configurations. Compared with usefulness, feasibility assessments exhibit higher and more stable agreement levels, indicating greater consistency among domain-specific evaluators. The AI with RAG and AI without RAG configurations demonstrate slight to fair agreement, reflecting moderate alignment in technical feasibility judgements. The Human configuration achieves moderate agreement at both pairwise and overall levels, suggesting more coherent and aligned feasibility assessments among experts. The Hybrid configuration shows fair agreement, indicating improved consistency relative to purely AI-based evaluations but lower alignment than the human panel. Overall, these results indicate that feasibility evaluations involve more differentiated and domain-sensitive reasoning, leading to more meaningful inter-rater consistency than usefulness assessments.

Table 5. Inter-rater agreement for feasibility evaluations

Configuration	Avg. Cohen's κ	Fleiss' κ	Interpretation
AI with RAG	0.167	0.147	Slight agreement in both pairwise and overall measures
AI without RAG	0.253	0.112	Fair pairwise agreement and slight overall agreement
Human expert	0.517	0.460	Moderate agreement in both pairwise and overall measures
Hybrid	0.300	0.223	Fair agreement in both pairwise and overall measures

Extending the analysis to sustainability, Table 6 reveals a further reduction in agreement levels. The AI with RAG configuration exhibits near-constant ratings, resulting in undefined pairwise agreement and negative overall kappa values, which indicate limited measurable consistency. The AI without RAG configuration demonstrates slight agreement, suggesting modest alignment among evaluators in

https://doi.org/10.1017/pds.2026.10603 Published online by Cambridge University Press

sustainability assessments. In contrast, the Human and Hybrid configurations exhibit negative pairwise and overall kappa values, indicating agreement below chance level and substantial variability in domain-specific judgements. Overall, the results suggest that sustainability evaluations are characterised by greater divergence among evaluators, reflecting the complexity and multidimensional nature of sustainability considerations in early-stage design.

Table 6. Inter-rater agreement for sustainability evaluations

Configuration	Avg. Cohen's κ	Fleiss' κ	Interpretation
AI with RAG	–	−0.227	Near-constant ratings with no overall agreement
AI without RAG	0.119	0.120	Slight agreement in both pairwise and overall measures
Human expert	−0.267	−0.227	No agreement in both pairwise and overall measures
Hybrid	−0.267	−0.500	No agreement in both pairwise and overall measures

4.2. Influence of retrieval-augmented knowledge (RQ2)

Table 7 compares the panel consensus scores for usefulness between AI configurations with and without retrieval-augmented knowledge. The AI without RAG configuration produces identical mean scores across all solution concepts, indicating highly uniform and stable evaluations. This suggests that, in the absence of external knowledge support, the panel tends to converge on similar usefulness judgements across different solution concepts. In contrast, the AI with RAG configuration achieves a slightly higher mean score with low variability, reflecting a modest improvement in perceived usefulness while maintaining stable evaluation behaviour. Overall, these results indicate that retrieval-augmented knowledge marginally increases the mean usefulness ratings without introducing substantial instability.

Table 7. Panel consensus comparison between AI configurations for usefulness evaluation

	Solution concept (SC)				
Configuration	SC1	SC2	SC3	Mean	SD
AI without RAG	3.67	3.67	3.67	3.67	0.00
AI with RAG	3.83	3.67	3.67	3.72	0.10

Following the analysis of usefulness, Table 8 examines the effect of retrieval-augmented knowledge on feasibility evaluations. The AI without RAG configuration achieves a slightly higher overall mean score but exhibits greater variability across solution concepts, indicating less stable feasibility assessments. This suggests that, in the absence of retrieved knowledge, feasibility judgements are more sensitive to contextual differences. By contrast, the AI with RAG configuration demonstrates a marginally lower mean score accompanied by reduced variability, reflecting more consistent evaluation behaviour. The lower standard deviation indicates that retrieval-augmented knowledge contributes to greater stability in feasibility assessments, even though it does not substantially increase perceived feasibility. Overall, these findings suggest that RAG appears to contribute to more stable feasibility evaluations.

Table 8. Panel consensus comparison between AI configurations for feasibility evaluation

	Solution concept (SC)				
Configuration	SC1	SC2	SC3	Mean	SD
AI without RAG	2.67	3.33	3.00	3.00	0.33
AI with RAG	2.67	3.00	3.00	2.89	0.19

Building on the analysis of usefulness and feasibility, Table 9 examines the influence of retrieval-augmented knowledge on sustainability evaluations. The AI without RAG configuration achieves a slightly higher overall mean score with moderate stability, indicating relatively consistent sustainability assessments across solution concepts. The AI with RAG configuration produces a comparable mean score and slightly lower variability, reflecting similarly stable evaluation behaviour. The small difference in mean values suggests that retrieval-augmented knowledge does not substantially alter the perceived

https://doi.org/10.1017/pds.2026.10603 Published online by Cambridge University Press

sustainability of the solution concepts. However, the marginal reduction in variability indicates that RAG contributes to more uniform sustainability judgements. Overall, these findings imply that the primary impact of RAG in sustainability evaluation lies in enhancing consistency rather than substantially altering overall panel ratings.

Table 9. Panel consensus comparison between AI configurations for sustainability evaluation

	Solution concept (SC)				
Configuration	**SC1**	**SC2**	**SC3**	**Mean**	**SD**
AI without RAG	3.00	3.33	3.33	3.22	0.19
AI with RAG	3.17	3.00	3.33	3.17	0.17

4.3. Comparative performance of AI, human, and hybrid panels (RQ3)

This section compares how closely the AI-based configurations match human expert judgement using Spearman rank correlation and mean absolute error (MAE). Table 10 summarises the results across the three solution concepts. The AI without RAG configuration shows moderate positive alignment with a slightly higher MAE, suggesting partial similarity in the overall evaluation patterns and assigned scores. In contrast, the AI with RAG setting shows a negative correlation and the largest MAE, indicating that access to retrieved information does not necessarily lead to human-like ranking behaviour in this experiment. Interestingly, the hybrid configuration demonstrates the closest alignment with the human panel in this experimental setting, achieving both the highest correlation and the lowest MAE. This suggests that combining human input with AI support produces evaluations that more closely reflect human judgement in this experimental setting. Overall, the results for RQ3 show that while AI can assist in early-stage evaluation, the hybrid approach provides the closest approximation to human assessment patterns.

Table 10. Agreement and deviation between AI, hybrid and human evaluations

Configuration	**Spearman correlation vs Human**	**MAE vs Human**
AI without RAG	0.50	0.26
AI with RAG	−0.86	0.30
Hybrid	1.00	0.15

Taken together, the results indicate that the proposed framework is intended to structure early-stage evaluation rather than replace expert judgement. It is particularly relevant in exploratory screening contexts, where multiple alternatives must be assessed systematically under time constraints. The framework contributes by organising multi-perspective evaluation and making underlying reasoning more transparent. However, its effectiveness depends on the relevance and coverage of retrieved knowledge, and it does not substitute for experiential insight or organisational considerations that shape real-world decisions. Variability among experts should therefore not be interpreted solely as a weakness. In early design phases, differences in judgement may reflect uncertainty, emerging risks or tacit knowledge. In this sense, the hybrid configuration can support more reflective and robust early-stage decision-making.

4.4. Limitations and future work

This study has several limitations that should be acknowledged. The experiment involved a small number of solution concepts and relied on a specific set of knowledge sources, which limits statistical robustness and may not represent broader engineering contexts. In particular, the rank correlation analysis was based on three solution concepts, which restricts the statistical generalisability of the alignment results. Although demonstrated within a process engineering case, the framework is conceptually transferable to other domains requiring structured multi-criteria evaluation. The impact of the RAG configuration depended on the relevance of retrieved documents and may vary when sources are incomplete or misaligned with the problem context. Furthermore, the evaluation employed a simplified consensus mechanism and focused on usefulness, technical feasibility and environmental sustainability, without incorporating additional factors, such as organisational, economic or strategic considerations that

https://doi.org/10.1017/pds.2026.10603 Published online by Cambridge University Press

frequently shape real-world decision-making. While GPT-5.1 was used in this study, the framework is model-agnostic and can be implemented with other LLMs, subject to comparable reasoning capabilities. Future work should expand broader concept sets, domain-specific knowledge bases and richer human-AI collaboration formats.

5. Conclusion

This study examined a multi-agent framework for early-stage concept evaluation, focusing on evaluation consistency, the influence of retrieval-augmented knowledge and alignment with human expert judgement. The findings show that AI-based panels tend to produce uniform evaluation patterns across solution concepts, though statistical agreement levels vary depending on the evaluation criterion. The inclusion of retrieval-augmented knowledge alters evaluation patterns but does not lead to closer correspondence with human rating behaviour. Across all configurations, the hybrid setting achieved the closest alignment with human judgement, indicating that combining AI-supported assessment with expert interpretation provides the most stable outcomes. Overall, the findings indicate that while AI can support early design evaluation, human involvement remains essential. The proposed framework offers a practical basis for structured, knowledge-supported assessment, and future studies may further enhance its effectiveness through broader test cases and more comprehensive knowledge integration.

Acknowledgement

The authors are grateful to the German Academic Exchange Service (DAAD) for funding this research.

References

Baffo, I., Leonardi, M., Bossone, B., Camarda, M. E., D'Alberti, V., & Travaglioni, M. (2023). A decision support system for measuring and evaluating solutions for sustainable development. *Sustainable Futures*, 5, 100109. https://doi.org/10.1016/j.sftr.2023.100109

Blessing, L. T. M., & Chakrabarti, A. (2009). Descriptive Study II: Evaluating Design Support. In *DRM, a Design Research Methodology* (pp. 181–214). Springer London. https://doi.org/10.1007/978-1-84882-587-1_6

Camacho, L., Dumée, L., Zhang, J., *et al.* (2013). Advances in Membrane Distillation for Water Desalination and Purification Applications. *Water*, 5(1), 94–196. https://doi.org/10.3390/w5010094

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, 7(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

Chen, Z., Zhong, P., Liu, M., Ma, Q., & Si, G. (2022). An integrated expert weight determination method for design concept evaluation. *Scientific Reports*, 12(1), 6358. https://doi.org/10.1038/s41598-022-10333-6

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Dahiru Umar, L., Azeem, M. A., Baroud, T. N., & Al Abdulgader, H. (2022). *Permeate gap membrane distillation* (Patent US11345619B1). USPTO.

Dean, D., Hender, J., Rodgers, T., & Santanen, E. (2006). Identifying Quality, Novel, and Creative Ideas: Constructs and Scales for Idea Evaluation. *Journal of the Association for Information Systems*, 7(10), 646–699. https://doi.org/10.17705/1jais.00106

Drioli, E. (2017). *Membrane Distillation*. MDPI. https://doi.org/10.3390/books978-3-03842-461-1

Dym, C. L., Little, P., & Orwin, E. (2013). *Engineering Design: A Project-Based Introduction* (4th ed.). John Wiley & Sons.

Edwards, K. M., Tehranchi, F., Miller, S. R., & Ahmed, F. (2025). *AI Judges in Design: Statistical Perspectives on Achieving Human Expert Equivalence With Vision-Language Models*. https://arxiv.org/pdf/2504.00938

Elliott, K., Potts, E., & Bercik, I. (2025). *SRM NIST Standard Reference Materials catalog January 2025* (NIST Special Publication; 260-176-2025). U.S. Dept. of Commerce, National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.260-176-2025

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382. https://doi.org/10.1037/H0031619

Francis, L., Ahmed, F. E., & Hilal, N. (2022). Advances in Membrane Distillation Module Configurations. *Membranes*, 12(1), 81. https://doi.org/10.3390/membranes12010081

Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). *Large Language Model based Multi-Agents: A Survey of Progress and Challenges*. http://arxiv.org/abs/2402.01680

ISO 14040. (2006). *Environmental management - Life cycle assessment - Principles and framework.* https://www.iso.org/standard/37456.html

https://doi.org/10.1017/pds.2026.10603 Published online by Cambridge University Press

Klüppel, H.-J. (2005). The Revision of ISO Standards 14040-3 - ISO 14040: Environmental management – Life cycle assessment – Principles and framework - ISO 14044: Environmental management – Life cycle assessment – Requirements and guidelines. *The International Journal of Life Cycle Assessment*, 10(3), 165–165. https://doi.org/10.1065/lca2005.03.001

Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159. https://doi.org/10.2307/2529310

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2021). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, *2020-December*. http://arxiv.org/abs/2005.11401

Livotov, P., Huber, N., Hentschel, C., *et al.* (2026). Comparative Study on AI-Augmented Inventive Problem Solving with TRIZ, TRIZ-AI Hybrid, and Autonomous AI: An Inline Coating Measurement Case in Lithium-Ion Cell Production. In D. Cavallucci, S. Brad, P. Livotov, & R. Houssin (Eds.), *World Conference of AI-Powered Innovation and TRIZ Methodology* (pp. 35–54). Springer Nature Switzerland.

Livotov, P., & Mas'udah, M. (2025). AI-powered inventive design: idea funnelling, concept creation, and hybrid problem-solving teams. *Proceedings of the Design Society*, 5, 479–488. https://doi.org/10.1017/pds.2025.10062

Mas'udah, Livotov, P., Hartmann, N., *et al.* (2026). Enhancing AI-Generated Solutions with Visual Retrieval-Augmented Generation (V-RAG): A Framework for Problem Formulation. In D. Cavallucci, S. Brad, P. Livotov, & R. Houssin (Eds.), *World Conference of AI-Powered Innovation and TRIZ Methodology* (pp. 21–36). Springer Nature Switzerland.

Mas'udah, Livotov, P., & Nugroho, S. (2024). Evaluating AI-generated solution ideas: A comparative study of AI and human assessments for sustainable process design. *Proceedings of the Upper-Rhine Artificial Intelligence Symposium*. https://doi.org/10.60643/urai.v2024p113

Mas'udah, M., Livotov, P., Hartmann, N., *et al.* (2025). Optimising AI-driven solutions without trade-offs: predicting and preventing potential failures in sustainable innovation. *Proceedings of the Design Society*, 5, 2561–2570. https://doi.org/10.1017/pds.2025.10270

Mesbah, S., Arous, I., Yang, J., & Bozzon, A. (2023). HybridEval: A Human-AI Collaborative Approach for Evaluating Design Ideas at Scale. *Proceedings of the ACM Web Conference 2023*, 3837–3848. https://doi.org/10.1145/3543507.3583496

OpenAI. (2025, October 24). *ChatGPT homepage*. https://chat.openai.com

Pahl, G., Beitz, W., Feldhusen, J., & Grote, K.-H. (2007). Product Planning, Solution Finding and Evaluation. In *Engineering Design* (pp. 63–124). Springer London. https://doi.org/10.1007/978-1-84628-319-2_3

Qian, Y., Zhang, S., Zhou, Y., Ding, H., Socolinsky, D., & Zhang, Y. (2025). Enhancing LLM-as-a-judge via multi-agent collaboration. *WMAC 2025 : AAAI 2025 Workshop on Advancing LLM-Based Multi-Agent Collaboration*. https://www.amazon.science/publications/enhancing-llm-as-a-judge-via-multi-agent-collaboration

Rasheed, Z., Waseem, M., Systä, K., & Abrahamsson, P. (2024). *Large Language Model Evaluation Via Multi AI Agents: Preliminary results*. http://arxiv.org/abs/2404.01023

Salinas-Rodríguez, S. G., & Schippers, J. C. (2021). *Introduction to desalination. In Seawater Reverse Osmosis Desalination: Assessment and Pre-treatment of Fouling and Scaling* (pp. 1–27). IWA Publishing. https://doi.org/10.2166/9781780409863_0001

Sauer, C. R., & Burggräf, P. (2025). Hybrid intelligence – systematic approach and framework to determine the level of Human-AI collaboration for production management use cases. *Production Engineering*, 19(3–4), 525–541. https://doi.org/10.1007/s11740-024-01326-7

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1), 72. https://doi.org/10.2307/1412159

Ullman, D. G. (2003). *The Mechanical Design Process* (3rd ed.). McGraw-Hill.

UNEP. (2011). Life Cycle Sustainability Assessment. https://www.lifecycleinitiative.org/starting-life-cycle-thinking/life-cycle-approaches/life-cycle-sustainability-assessment/

United Nations. (2015). *Transforming our world: the 2030 Agenda for Sustainable Development*. https://sdgs.un.org/2030agenda

Wail Sulaiman, F., & Lawal, D. U. (2022). *Bubble feed membrane distillation system (Patent US11413581B2)*. USPTO.

Wu, Q., Bansal, G., Zhang, J., *et al.* (2023). *AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation*. http://arxiv.org/abs/2308.08155

Xu, W., Kotecha, M. C., & McAdams, D. A. (2024). How good is ChatGPT? An exploratory study on ChatGPT's performance in engineering design tasks and subjective decision-making. *Proceedings of the Design Society*, 4, 2307–2316. https://doi.org/10.1017/PDS.2024.233

https://doi.org/10.1017/pds.2026.10603 Published online by Cambridge University Press