EVALUATING AI-GENERATED SOLUTION IDEAS: A COMPARATIVE STUDY OF AI AND HUMAN ASSESSMENTS FOR SUSTAINABLE PROCESS DESIGN

Mas'udah, Pavel Livotov, Saptadi Nugroho

Offenburg University of Applied Sciences, Germany
Albert Ludwig University of Freiburg, Germany

OBJECTIVE

- To assess the ability of GPT-4o in autonomously evaluating its generated solution ideas.
- To compare AI evaluations with human expert assessments on key criteria: novelty, feasibility, usefulness, and sustainability

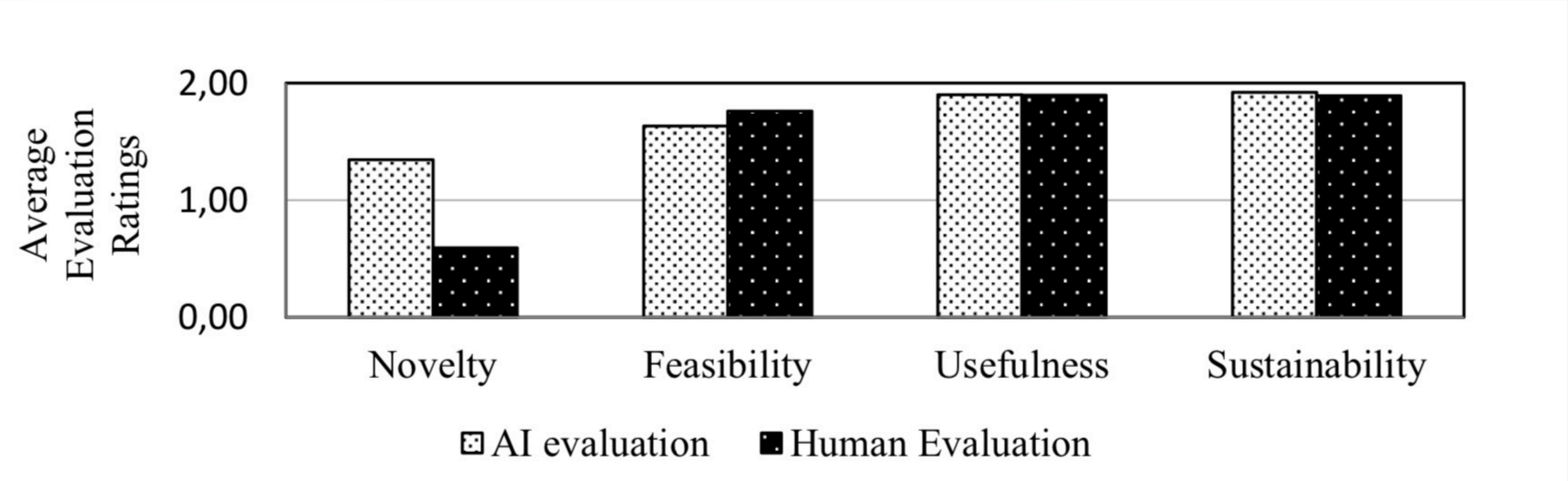
METHODOLOGY

- **Research Design**: A dual approach where GPT-4o was used for generating and evaluating solution ideas.
- **Case Study**: Froth flotation for nickel recovery, focusing on sustainability and reduced chemical use.
- **Evaluation Metrics**: Assessment based on novelty, feasibility, usefulness, and sustainability.
- **Evaluation Process**: 50 AI-generated ideas were rated by GPT-4o and two human experts, with scores compared using Cohen's and Fleiss' Kappa for inter-rater reliability.

RESULTS

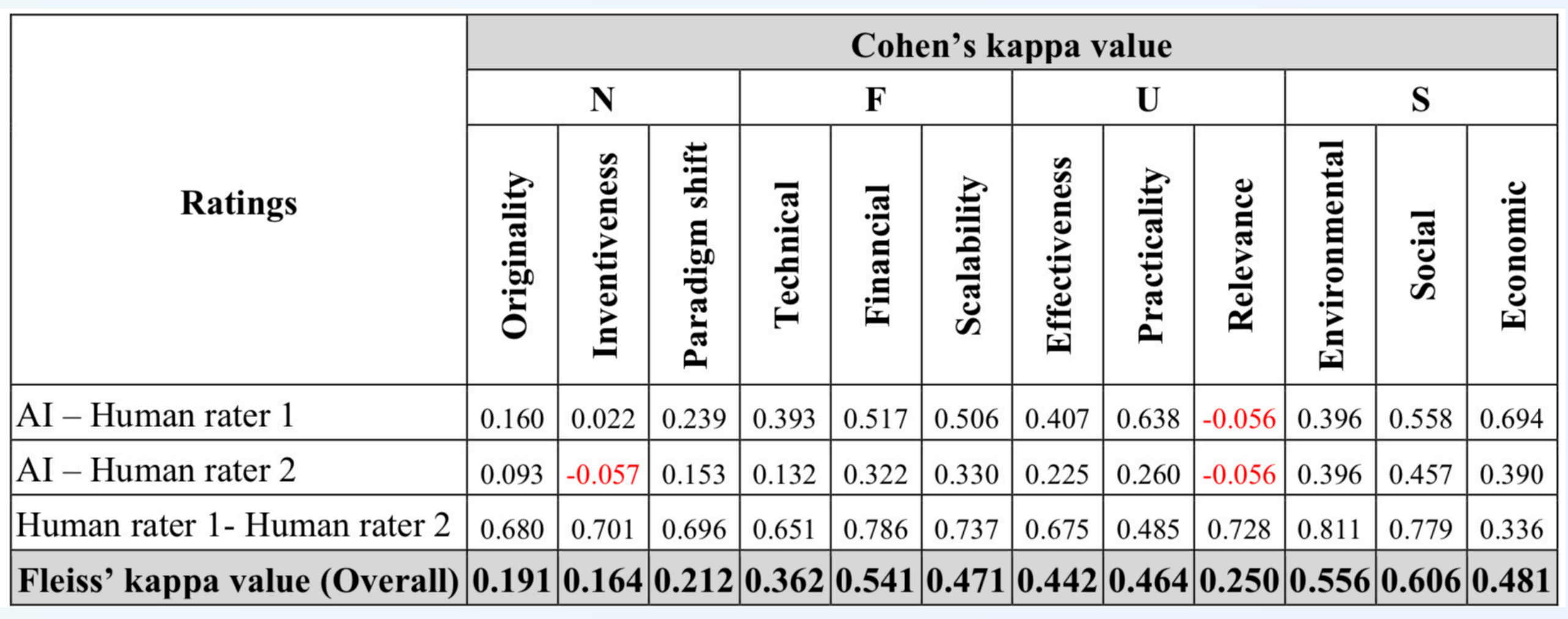
- Strong alignment between AI and human evaluations for feasibility, usefulness, and sustainability.
- AI perceived higher novelty scores than humans, indicating differences in criteria interpretation.
- Higher agreement in environmental and social sustainability metrics, but lower in novelty

EVALUATION RATINGS



- **Strong alignment** between AI and human assessments in feasibility, usefulness, and sustainability, indicating AI's potential for effective preliminary evaluations.
- **Notable discrepancy in novelty**: AI tends to rate ideas as more original compared to human experts, suggesting differences in interpretation and stricter human standards.
- **AI's strength** lies in assessing feasibility, usefulness, and sustainability, while human insight is essential for evaluating novelty and originality.

INTER-RATER AGREEMENT



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**

- **Highest agreement observed in sustainability**: Cohen's and Fleiss' Kappa values indicate strong consistency in ratings, especially in environmental and social aspects.
- **Moderate agreement in feasibility and usefulness**, showing that AI assessments are reliable but may vary in subjective interpretations.
- **Lower agreement in novelty**, highlighting the challenge AI faces in matching human evaluations on originality and inventiveness.
- **Overall, the agreement indicates** that while AI aligns well with human evaluations for feasibility, usefulness, and sustainability, novelty remains an area requiring further refinement for better alignment.

CONCLUSION

- GPT-4o can serve as a preliminary evaluation tool with alignment in most criteria, though human expertise is essential for novelty assessments.
- A hybrid approach integrating AI and human insights provides a comprehensive evaluation framework.

FUTURE WORK

- Extend studies to different AI models and multiple case studies.
- Develop AI tools with training on creativity-specific datasets to improve novelty evaluations.
- Incorporate broader panels of human experts for more diverse comparison.